The Geometry of Semi-Supervised Learning

A THESIS PRESENTED
BY
LUKE MELAS-KYRIAZI
TO
THE DEPARTMENT OF MATHEMATICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
MATHEMATICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2020

Contents

1 Introduction 1

1.1 What is Learning? 1

1.2 Lessons from Human and Animal Learning 2

1.3 Types of Learning 3

1.4 Why Semi-Supervised Learning? 5

1.5 Overview 8

2 Foundations 9

2.1 The Learning Problem 10

2.2 Regularization 12

3 Kernel Learning 15

3.1 Reproducing Kernel Hilbert Spaces 17

3.2 Kernels, Positive Functions, and Feature Maps 27

3.3 Tikhonov Regularization and the Representer Theorem . . . 32

3.4 Algorithms 34

4 Graphs and Manifolds 40

4.1 Smoothness and the Laplacian 41

4.2 Lessons from Physics 50

4.3 The Laplacian Spectrum 55

4.4 Eigenvalues and Connectivity 64

4.5 The Heat Kernel 71

5 Manifold Regularization **75**

5.1 Manifold Regularization 76

5.2 Representer Theorems 78

5.3 Algorithms 86

5.4 Convergence of the Graph Laplacian 90

A Appendix **97**

A.1 Additional Proofs from Chapter 2 97

A.2 Additional Proofs from Chapter 4 101

A.3 Additional Proofs from Chapter 5 119

References **128**

List of figures

1.3.1 Types of Learning 4
1.4.1 Motivation for Semi-Supervised Learning 6

3.2.1 Kernels, Positive Functions, and Feature Maps 27

4.1.1 Classic Graphs 47
4.1.2 Illustration of Torus 48
4.2.1 Illustration of Tutte's Theorem 53
4.3.1 Isospectral Domains and Graphs 56
4.3.2 Eigenfunctions of the Circle and the Cycle Graph 59
4.4.1 Smooth Approximation to the Indicator 71

Acknowledgments

I wish to thank my family, my friends, and my wonderful advisor Professor Arjun Manrai for their support throughout this process.

1
Introduction

1.1 What is Learning?

From an early age, our parents and teachers impress upon us the importance of learning. We go to school, do homework, and write senior theses in the name of learning. But what exactly is learning?

Theories of learning, which aim to answer this question, stretch back as far as Plato. Plato's theory, as presented in the *Phaedo*, understands learning as the rediscovery of innate knowledge acquired at or before birth. For the past two millennia, epistemologists have debated the meaning and mechanisms of learning, with John Locke notably proposing a theory based on the passive acquisition of simple ideas. Scientific approaches to understanding learning emerged beginning in the nineteenth century. Ivan Pavlov's famous classical conditioning experiments, for example, demonstrated how dogs learned to associate one stimulus (i.e. ringing bells) with another (i.e. food). A multitude of disciplines now have subfields dedicated to theories of learning: psychology, neuroscience, pedagogy, and linguistics, to name only a few.

Over the past few decades, the rise and proliferation of computers has prompted researchers to consider what it means for a computer algorithm to learn. Specifically, the past two decades have seen a proliferation of research in machine learning, the study of algorithms that can perform tasks without being explicitly programmed. Now ubiquitous, these machine learning algorithms are integrated into a plethora of real-world systems and applications. From Google Search to Netflix's recommendation engine to Apple's Face ID software, much of the "intelligence" of modern computer applications is a product of machine learning.

This thesis takes a mathematical approach to machine learning, with the goal of building and analyzing theoretically-grounded learning algorithms. We focus in particular on the subfield of *semi-supervised learning*, in which machine learning models are trained on both unlabeled and labeled data. In order to understand modern semi-supervised learning methods, we develop an toolkit of mathematical methods in spectral graph theory and Riemannian geometry. Throughout the thesis, we will find that understanding the underlying mathematical structure of machine learning algorithms enables us to interpret, improve, and extend upon them.

1.2 Lessons from Human and Animal Learning

Although this thesis is concerned entirely with machine learning, the ideas presented within are grounded in our intuition from human and animal learning. That is, we design our mathematical models to match our intuition about what should and should not be considered learning.

An example here is illustrative. Consider a student who studies for a test using a copy of an old exam. If the student studies in such a way that he or she develops an understanding of the material and can answer new questions about it, he or she has learned something. If instead the student memorizes all the old exam's questions and answers, but cannot answer any new questions about the material, the student has not actually learned anything. In the jargon of machine learning, we would say that the latter student does not *generalize*: he makes few errors on the questions he has seen before (the *training* data) and many errors on the questions he has not seen before (the *test* data).

Our formal definition of learning, given in Chapter 2, will hinge upon this idea of generalization. Given a finite number of examples from which to learn, we would like to be able to make good predictions on new, unseen examples.

Our ability to learn from finite data rests on the foundational assumption that our data has some inherent structure. Intuitively, if we did not assume that our world had any structure, we would not be able to learn anything from past experiences; we need some prior knowledge, an *inductive bias*, to be able to generalize from observed data to unseen data. We can formalize this intuitive notion in the No Free Lunch Theorem, proven in Chapter 2.

Throughout this thesis, we adopt the inductive bias that the functions we work with should be simple. At a high level, this bias is Occam's Razor: we prefer simpler explanations of our data to more complex ones. Concretely, this bias takes the form of *regularization*, in which we enforce that the norm of our learned function is small.

The thesis builds up to a type of regularization called *manifold regularization*, in which the norm of our function measures its smoothness with respect to the manifold on which our data lie. Understanding manifold regularization requires developing a substantial amount of mathematical machinery, but it is worth the effort because it will enable us to express the inductive bias that our functions should be simple.

1.3 Types of Learning

In computational learning, types of learning are generally categorized by the data available to the learner. Below, we give an overview of the three primary types of computational learning: supervised, semi-supervised, and unsupervised learning. An illustration is shown in Figure 1.3.1.

1.3.1 Supervised Learning

The goal of supervised learning is to approximate a function $f : X \to Y$ using a training set $S = \{x_i, y_i\}_{i=1}^N$. Note that the space of inputs X and the space of outputs Y are entirely general. For example, X or Y may contain vectors, strings, graphs, or molecules. Usually, we will consider problems for which Y is $\mathbb{R}$ (regression) or for which Y is a set of classes $Y = \mathcal{C} = \{0, 1, \cdots, n-1\}$ (classification). The special case $Y = \{0, 1\}$ is called binary classification.

The defining feature of supervised learning is that the training set S is fully-labeled, which means that every point x_i has a corresponding label y_i.

Example: Image Classification Image classification is the canonical example of supervised learning in computer vision. Here, X is the set of (natural) images and Y is a set of $|\mathcal{C}|$ categories. Given an image $x_i \in X$, the task is to classify the image, which is to assign it a label $y_i \in Y$. The

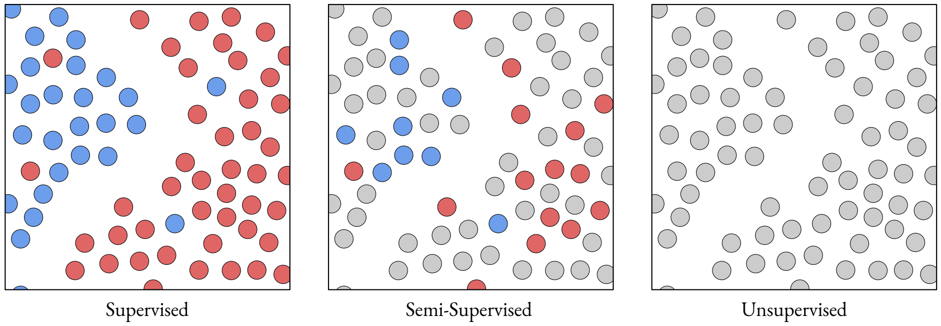


Figure 1.3.1: An illustration of supervised, semi-supervised, and unsupervised learning.

standard large-scale classification dataset ImageNet [26] has $|\mathcal{C}| = 1000$ categories and $|S| \approx 1,200,000$ labeled training images.

1.3.2 Semi-Supervised Learning

In semi-supervised learning, the learner is given access to labeled training set $S_L = \{x_i, y_i\}_{i=1}^{N_L}$ along with unlabeled data $S_U = \{x_i\}_{i=1}^{N_U}$. Usually, the size of the unlabeled data is much larger than the size of the labeled data: $N_U \gg N_L$.

It is possible to turn any semi-supervised learning problem into a supervised learning problem by discarding the unlabeled data S_U and training a model using only the labeled data S_L. The challenge of semi-supervised learning is to use the information in the unlabeled data to train a better model than could be trained with only S_L. Semi-supervised learning is the focus of this thesis.

Example: Semi-Supervised Node Classification In social networks, recommender systems, and many other domains, it is common to be given a graph $G = (V, E)$ where a subset of the $|V|$ nodes are labeled. For example, we may have a social network in which a small number of users have rated a certain movie, and we would like to predict how other users will rate the movie. In this case, $X = V$ and Y is the space of ratings (e.g. 1 to 5 stars). Using a semi-supervised learning algorithm, we can incorporate into our training process information about both users who rated movies (labeled data) and those who have not yet rated movies (unlabeled data).

1.3.3 Unsupervised Learning

In unsupervised learning, we are given data $X = \{x_i\}_{i=1}^{N}$ without any labels. In this case, rather than trying to learn a function f to a space of labels, we aim to learn useful representations or properties of our data. For example, we may try to cluster our data into semantically meaningful groups, learn a generative model of our data, or perform dimensionality reduction on our data.

Example: Dimensionality Reduction for Single-cell RNA Data Researchers in biology performing single-cell RNA sequencing often seek to visualize high-dimensional sequencing data. That is, they aim to embed their high-dimensional data into a lower-dimensional space (e.g. the $2D$ plane) in such a way that it retains its high-dimensional structure. They may also want to cluster their data either before or after applying dimensionality reduction. Both of these tasks may be thought of as unsupervised learning problems, as their goal is to infer the structure of unlabeled data.

Finally, we should note that there are a plethora of other subfields and subclassifications of learning algorithms: reinforcement learning, active learning, online learning, multiple-instance learning, and more.[1] For our purposes, we are only concerned with the three types of learning above.

1.4 Why Semi-Supervised Learning?

This thesis focuses on semi-supervised learning, the setting in which we have limited access to labeled training data. Semi-supervised learning is of great interest because it is often easy to collect large amounts of data but difficult or extremely expensive to label this data. Unfortunately, in practice, most people deal with this issue by discarding their unlabeled data and working only with a small labeled subset.

The following toy example shows that simply discarding unlabeled data is dangerous, because unlabeled data can change our notion of what a "good" solution looks like.

Toy Example Figure 1.4.1 presents a simple dataset of points in the $2D$ plane, with two points having labels (shown as red and blue). Suppose we

[1] For an in-depth review of many of these fields, reader is encouraged to look at [66].

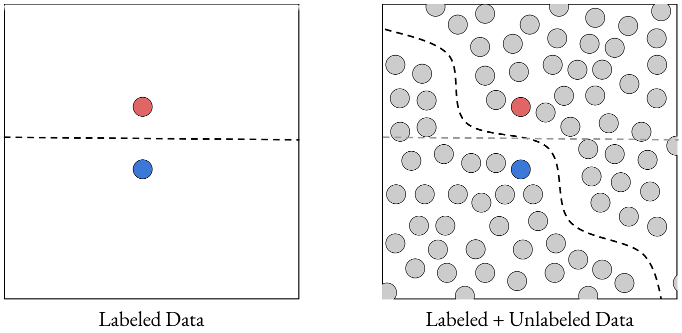


Figure 1.4.1: A toy example illustrating how unlabeled data can change how we see a binary classification problem. On the left, discarding the unlabeled data, the natural decision boundary between the red and blue classes is a straight line. On the right, with the unlabeled data, the natural decision boundary fits the contours of the unlabeled data.

wish to perform binary classification, which is to say separate the plane into two regions corresponding to the two classes.

If we discard our unlabeled data and look only at our two labeled data points, natural boundary between the two classes is a straight line separating the points. However, looking at both the unlabeled and labeled data, it is clear that this line between the two points is no longer a natural boundary. Instead, the natural boundary is a curved line conforming to the shape of the unlabeled data. In this way, the addition of unlabeled data has changed our notion of a good classification function.

This intuition extends to many real-world machine learning problems. Indeed, recent research has used unlabeled data to realize large performance gains on a wide range of problems. In particular, due to the ease of scraping unlabeled text, images, and videos from the internet, a growing number of problems involving these types of data are being approached as semi-supervised learning problems.

Example: Low-Resource Machine Translation Suppose we are tasked with building a translation system between two languages.

If these languages are commonly translated into one another, for example English and German, building a translation system is relatively straightforward: we collect a large dataset of documents that have already been translated into both languages by human experts, and then we train a model on this (paired) dataset.[2] This is a standard supervised learning problem, perhaps the most widely-studied supervised learning problem in the field of natural language processing (NLP).

If these languages are not commonly translated into one another, for example Finnish and Nepali, it may not be possible to find enough expertly translated documents to train a good machine learning model. However, we can easily collect a large amount of unpaired Finnish and Nepali text by scraping the internet. In total, we are able to collect a large amount of Finnish text, a large amount of Nepali text, and a small amount of paired Finnish-Nepali text. Here, semi-supervised learning algorithms enable us to utilize our unpaired texts to train a higher-quality translation model.

Example: Semantic Segmentation Another reason to use semi-supervised learning is that labeled data are sometimes very time-consuming or expensive to obtain. For example, consider semantic segmentation, the task of classifying every pixel of an image into one of a predefined set of semantic classes (e.g. tree, cars, birds, sky, etc.). An accurate semantic segmentation model is a key part of a self-driving car, as a self-driving car needs to identify where other vehicles, bikes, and pedestrians are on the road ahead of it.

Semantic segmentation is often tackled using neural network models, which require large numbers of labeled images to achieve good performance. However, labeling images for semantic segmentation is time-consuming and expensive. For example, on one dataset, each image took over 90 minutes for a human to annotate [24].[34] Self-supervised learning enables us to train a high-quality segmentation model using a small amount of this arduously-annotated labeled data and a large amount of easily-collected unlabeled data.

[2] In fact, this is how Google Translate works, as described in their paper here [89].

[3] Labeling images for segmentation is so arduous that it has become a large industry: Scale AI, a startup that sells data labeling services to self-driving car companies, is valued at over a billion dollars. According to their website, they charge $6.40 per annotated frame for image segmentation. If you were to record video at 30 frames-per-second for 24 hours and try to label every frame, you would have to label 2,592,000 images. Many of these images would be quite similar, but even if you subsampled to 1 frame-per-second, it would require labeling 86,400 images.

[4] Annotation is even more costly in domains such as medical image segmentation, where images must be annotated by highly-trained professionals.

1.5 Overview

This thesis presents a mathematical perspective on semi-supervised learning. Its primary goal is to give an exposition of *manifold regularization*, a form of regularization based on the spectral properties of a graph generated from the data. Whereas standard regularization techniques capture the idea that our function should be simple with respect to the ambient function space, manifold regularization captures the idea that our function should be simple with respect to the space on which the data are generated.

In order to understand manifold regularization, we first need to understand (1) kernel learning in a fully-supervised setting, and (2) the relationship between manifolds and graphs.

Chapters 2 and 3 are dedicated to (1). Chapter 2 lays the foundations for supervised and semi-supervised learning. Chapter 3 develops the theory of supervised kernel learning in Reproducing Kernel Hilbert Spaces. This theory lays mathematically rigorous foundations for large classes of regularization techniques.

Chapter 4 is dedicated to (2). It explores the relationship between graphs and manifolds through the lens of the Laplacian operator, a linear operator that can be defined on both graphs and manifolds. Although at first glance these two types of objects may not seem to be very similar, we will see that the Laplacian reveals a remarkable correspondence between them. By the end of the chapter, we will have developed a unifying mathematical view of these seemingly disparate techniques.

Finally, Chapter 5 presents manifold regularization. We will find that, using the Laplacian of a graph generated from our data, it is simple to add manifold regularization to many learning algorithms. At the end of the chapter, we will prove that this graph-based method is theoretically grounded: the Laplacian of the data graph converges to the Laplacian of the data manifold in the limit of infinite data.

This thesis is designed for a broad mathematical audience. Little background is necessary apart from a strong understanding of linear algebra (e.g. Math 25A). A few proofs will require additional background, such as familiarity with Riemannian geometry. Illustrative examples from mathematics and machine learning are incorporated into the text whenever possible.

2
Foundations

The first step in understanding machine learning algorithms is to define our learning problem. In this chapter, we will only work in the supervised setting, generally following the approaches from [19, 74, 76]. Chapter 5 will extend the framework developed here to a semi-supervised setting.

2.0.1 Learning Algorithms & Loss Functions

A learning algorithm $\mathcal{A}$ is a map from a finite dataset S to a candidate function $\hat{f}$, where $\hat{f}$ is measurable. Note that $\mathcal{A}$ is stochastic because the data S is a random variable. We assume that our data (x_i, y_i) are drawn independently and identically distributed from a probability space $X \times Y$ with measure ρ.

We define what it means to "do well" on a task by introducing a loss function, a measurable function $L : X \times Y \times F \to [0, \infty)$. This loss almost always takes the form $L(x, y, f) = L'(y, f(x))$ for some function L', so we will write the loss in this way moving foward. Intuitively, we should think of $L(y, \hat{f}(x))$ as measuring how costly it is to make a prediction $\hat{f}(x)$ if the true label for x is y. If we predict $f(x) = y$, which is to say our prediction at x is perfect, we would expect to incur no loss at x (i.e. $L(y, \hat{f}(x)) = 0$).

Choosing an appropriate loss function is an important part of using machine learning in practice. Below, we give examples of tasks with different data spaces X, Y and different loss functions L.

Example: Image Classification Image classification, the task of classifying an image x into one of C possible categories, is perhaps the most widely-studied problem in computer vision. Here $x \in R^{H \times W \times 3}$, where H and W are the image height and width, and 3 corresponds to the three color channels (red, green, and blue). Our label space is a finite set $Y = \mathcal{C}$ where $|\mathcal{C}| = C$. A classification model outputs a discrete distribution $f(x_i) = p = (p_1, \ldots, p_C)$ over classes, with p_c corresponding to the probability that the input image x has class c.

As our loss function, we use cross-entropy loss:

$$L(y, f(x)) = -\frac{1}{N} \sum_{c=1}^{C} \mathbf{1}\{y_i = c\} \log(p_c), \quad p = f(x_i)$$

Example: Semantic Segmentation As mentioned in the introduction, semantic segmentation is the task of classifying every pixel in an input image. Here, $X = \mathcal{C}^{H \times W \times 3}$ like in image classification above, but $Y = \mathcal{C}^{H \times W}$ unlike above. The output $f(x) = p = (p_c^{(h,w)})$ is a distribution over classes for each pixel.

As our loss function, we use cross-entopy loss averaged across pixels:

$$L(y, f(x)) = -\frac{1}{N \cdot H \cdot W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} 1\{y_i^{(h,w)} = c\} \log(p_c^{(h,w)}), \quad p = f(x_i^{(h,w)})$$

Example: Crystal Property Prediction A common task in materials science is to predict the properties of a crystal (e.g. formation energy) from its atomic structure (an undirected graph). As a learning problem, this is a regression problem with X as the set of undirected graphs and $Y = \mathbb{R}$.

For the loss function, it is common to use mean absolute error (MAE) due to its robustness to outliers:

$$L(y, f(x)) = |y - f(x)|$$

2.1 The Learning Problem

Learning is about finding a function $\hat{f}$ that generalizes from our finite data S to the infinite space $X \times Y$. This idea may be expressed as minimizing the expected loss $\mathcal{E}$, also called the risk:

$$\mathcal{E}(f) = \mathbb{E}[L(y, f(x))] = \int_{X \times Y} L(y, f(x))\, d\rho(x, y)$$

Our objective in learning is to minimize the risk:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}[L(y, f(x))] = \arg\min_{f \in \mathcal{F}} \int_{X \times Y} L(y, f(x))\, d\rho(x, y)$$

Since we have finite data, even computing the risk is impossible. Instead, we approximate it using our data, producing the empirical risk:

$$\hat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^{N} L(y_i, f(x_i)) \approx \int_{X \times Y} L(y, f(x))\, d\rho(x, y) \tag{2.1}$$

This concept, *empirical risk minimization*, is the basis of much of modern machine learning.

One might hope that by minimizing the empirical risk over all measurable functions, we would be able to approximate the term on the right hand side of 2.1 and find a function $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mathcal{E}}(f)$ resembling the desired function f^*. However, without additional assumptions or priors, this is not possible. In this unconstrained setting, no model can achieve low error across all data distributions, a result known as the No Free Lunch Theorem.

The difference between the performance of our empirically learned function $\hat{f}$ and the best possible function is called the generalization gap or generalization error. We aim to minimize the probability that this error exceeds ε:

$$\mathbb{P}\left(\mathcal{E}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) > \varepsilon\right)$$

Note that here $\mathbb{P}$ refers to the measure ρ^N and that $\hat{f}$ is a random variable because it is the output of A with random variable input S.[1]

It would be desirable if this gap were to shrink to zero in the limit of infinite data:

$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{E}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) > \varepsilon\right) = 0 \qquad \forall \varepsilon > 0 \tag{2.2}$$

[1] Technically $\mathcal{A}$ could also be random, but for simplicity we will only consider deterministic A and random S here.

A learning algorithm with this property is called *consistent* with respect to ρ. Stronger, if property 2.2 holds for all fixed distributions ρ, the algorithm is *universally consistent.* Even stronger still, an algorithm that is consistent across finite samples from all distributions is *uniformly universally consistent*:

$$\lim_{n\to\infty} \sup_{\rho} \mathbb{P}\left(\mathcal{E}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{E}(f) > \varepsilon \right) = 0 \qquad \forall \varepsilon > 0 \tag{2.3}$$

Unfortunately, this last condition is *too* strong. This is the famous "No Free Lunch" Theorem.

Theorem 2.1.1 (No Free Lunch Theorem)**.** *No learning algorithm achieves uniform universal consistency. That is, for all $\varepsilon > 0$:*

$$\lim_{n\to\infty} \sup_{\rho} \mathbb{P}\left(\mathcal{E}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{E}(f) > \varepsilon \right) = \infty$$

We include a proof of this result for binary classification in Appendix A.1.1.

2.2 Regularization

The No Free Lunch Theorem states that learning in an entirely unconstrained setting is impossible. Nonetheless, if we constrain our problem, we can make meaningful statements about our ability to learn.

Looking at Equation 2.3, there are two clear ways to constrain the learning problem: (1) restrict ourselves to a class of probability distributions, replacing $\sup_\rho$ with $\sup_{\rho\in\Theta}$, or (2) restrict ourselves to a limited class of target functions $\mathcal{H}$, replacing $\inf_{f\in\mathcal{F}}$ with $\inf_{f\in\mathcal{H}}$. We examine the latter approach, as is common in statistical learning theory.

To make learning tractable, we optimize over a restricted set of hypotheses $\mathcal{H}$. But how should we choose $\mathcal{H}$? On the one hand, we would like $\mathcal{H}$ to be large, so that we can learn complex functions. On the other hand, with large $\mathcal{H}$, we will find complex functions that fit our training data but do not generalize to new data, a concept known as *overfitting.*

Ideally, we would like to be able to learn complex functions when we have a lot of data, but prefer simpler functions to more complex ones when we have little data. We introduce *regularization* for precisely this purpose. Regularization takes the form of a penalty R added to our loss term, biasing learning toward simpler and smoother functions.

Most of this thesis is concerned with the question of what it means to be a "simple" or "smooth" function. Once we can express and compute what it means to be simple or smooth, we can add this as a regularization term to our loss.

Moreover, if we have any tasks or problem-specific notions of what it means to be a simple function, we can incorporate them into our learning setup as regularization terms. In this way, we can inject into our algorithm prior knowledge about the problem's structure, enabling more effective learning from smaller datasets.

With regularization, learning problem turns into:

$$\arg\min_{f \in \mathcal{H}} \hat{\mathcal{E}}(f, x, y) + \lambda R(f, x, y)$$

where $\mathcal{H}$ can be a relatively large hypothesis space.

The parameter λ balances our empirical risk term and our regularization term. When λ is large, the objective is dominated by the regularization term, meaning that simple functions are preferred over ones that better fit the data. When λ is small, the objective is dominated by the empirical risk term, so functions with lower empirical risk are preferred even when they are complex. Tuning λ is an important element of many practical machine learning problems, and there is a large literature around automatic selection of λ [2].

Notation: The full expression $L + \lambda R$ is often called the loss function and denoted by the letter L. We will clarify notation in the following chapters whenever it may be ambiguous.

Often, R depends only on the function f and its parameters. We will call this data-independent regularization and write $R(f)$ for ease of notation. The reader may be familiar with common regularization functions (e.g. L1/L2 weight penalties), nearly all of which are data-independent. Manifold regularization, explored in Chapter 5, is an example of data-dependent regularization.

Example (Data-Independent): Linear Regression In linear regression, it is common to add a regularization term based on the magnitude of the weights to the standard least-squares objective:

$$R(f) = ||w||^\alpha \text{ for } \alpha > 0$$

When $\alpha = 2$, this is denoted Ridge Regression, and when $\alpha = 1$, it is denoted Lasso Regression. Both of these are instances of Tikhonov regularization, a data-independent regularization method explored in the following chapter.

Example (Data-Dependent): Image Classification When dealing with specialized domains such as images, we can incorporate additional inductive biases into our regularization framework. For example, we would expect an image to be classified in the same category regardless of whether it is rotated slightly, cropped, or flipped along a vertical line.

Recent work in visual representation learning employs these transformations to define new regularization functions. For example, [91] introduces a regularization term penalizing the difference between a function's predictions on an image and an augmented version of the same image:

$$R(f, x) = KL(f(x), f(\text{Aug}(x))$$

where Aug is an augmentation function, such as rotation by 15°, and $KL(\cdot, \cdot)$ is the Kullback–Leibler divergence, a measure of the distance between two distributions (because $f(x)$ is a distribution over C possible classes). This method currently gives state-of-the-art performance on image classification in settings with small amounts of labeled data [91].

3
Kernel Learning

In the previous chapter, we described the learning problem as the minimization of the regularized empirical risk over a space of functions $\mathcal{H}$.

This chapter is dedicated to constructing an appropriate class of function spaces $\mathcal{H}$, known as *Reproducing Kernel Hilbert Spaces*. Our approach is inspired by [12, 62, 68, 74].

Once we understand these spaces, we will find that our empirical risk minimization problem can be greatly simplified. Specifically, the Representer Theorem 3.3.1 states that its solution can be written as the linear combination of functions (kernels) evaluated at our data points, making optimization over $\mathcal{H}$ as simple as optimization over $\mathbb{R}^n$.

At the end of the chapter, we develop these tools into the general framework of *kernel learning* and describe three classical kernel learning algorithms. Due to its versatility and simplicity, kernel learning ranks among the most popular approaches to machine learning in practice today.

3.0.1 Motivation

Our learning problem, as developed in the last chapter, is to minimize the regularized empirical risk

$$\arg\min_{f \in \mathcal{H}} \hat{\mathcal{E}}(f, x, y) + \lambda R(f, x, y)$$

over a hypothesis space $\mathcal{H}$. The regularization function R corresponds to the inductive bias that simple functions are preferable to complex ones, effectively enabling us to optimize over a large space $\mathcal{H}$.

At this point, two issues remain unresolved: (1) how to define $\mathcal{H}$ to make optimization possible, and (2) how to define R to capture the complexity of a function.

If our functions were instead vectors in $\mathbb{R}^d$, both of our issues would be immediately resolved. First, we are computationally adept at solving optimization problems over finite-dimensional Euclidean space. Second, the linear structure of Euclidean space affords us a natural way of measuring the size or complexity of vectors, namely the norm $\|v\|$. Additionally, over the course of many decades, statisticians have developed an extensive theory of linear statistical learning in $\mathbb{R}^d$.

In an ideal world, we would be able to work with functions in $\mathcal{H}$ in the same way that we work with vectors in $\mathbb{R}^d$. It is with this motivation that mathematicians developed Reproducing Kernel Hilbert Spaces.

Informally, a Reproducing Kernel Hilbert Space (RKHS) is a potentially-infinite-dimensional space that looks and feels like Euclidean space. It is defined as a Hilbert space (a complete inner product space) satisfying an additional smoothness property (the *reproducing* property). Like in Euclidean space, we can use the norm $\|\cdot\|_K$ corresponding to the inner product of the RKHS to measure the complexity of functions in the space. Unlike in Euclidean space, we need an additional property to ensure that if two functions are close in norm, they are also close pointwise. This property is essential because it ensures that functions with small norm are near 0 everywhere, which is to say that there are no "complex" functions with small norm.

An RKHS is associated with a kernel $K : X \times X \to \mathbb{R}$, which may be thought of as a measure of the similarity between two data points $x, x' \in X$. The defining feature of kernel learning algorithms, or optimization problems over RKHSs, is that the algorithms access the data *only* by means of the kernel function. As a result, kernel learning algorithms are highly versatile; the data space X can be anything, so long as one can define a similarity measure between pairs of points. For example, it is easy to construct kernel learning algorithms for molecules, strings of text, or images.

3.1 Reproducing Kernel Hilbert Spaces

We are now ready to formally introduce Reproducing Kernel Hilbert Spaces.

Recall that a *Hilbert space* V is a complete vector space equipped with an inner product $\langle \cdot, \cdot \rangle$. In this chapter (except for a handful of examples), we will only work with real vector spaces, but all results can be extended without much hassle to complex-dimensional vector spaces.

For a set X, we denote by $\mathbb{R}^X$ the set of functions $X \mapsto \mathbb{R}$. We give $\mathbb{R}^X$ a vector space structure by defining addition and scalar multiplication pointwise:

$$(f_1 + f_2)(x) = f_1(x) + f_2(x) \qquad (a \cdot f)(x) = a \cdot f(x)$$

Linear functionals, defined as members of the dual space of R^X, may be thought of as linear functions $\mathbb{R}^X \to \mathbb{R}$. A special linear functional e_x, called the *evaluation functional*, sends a function f to its value at a point x:

$$e_x(f) = f(x)$$

When these evaluation functionals are bounded, our set takes on a remarkable amount of structure.

Definition 3.1.1 (RKHS)**.** Let X be a nonempty set. We say $\mathcal{H}$ is a *Reproducing Kernel Hilbert Space* on X if

1. $\mathcal{H}$ is a vector subspace of $\mathbb{R}^X$
2. $\mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle$ (it is a Hilbert Space)
3. For all $x \in X$, the linear evaluation functional $e_x : \mathcal{H} \to \mathbb{R}$ is bounded.

The last condition implies that e_x is continuous (even Lipschitz continuous). To see this, we can write:

$$\|e_x(f+h) - e_x(f)\| = \|e_x(h)\| \leq M \|h\| \qquad \text{for some constant } M$$

Letting $\|h\| \to 0$, we have the continuity of e_x.

Importantly, by the well-known Riesz Representation Theorem, each evaluation functional $e_x : \mathcal{H} \to \mathbb{R}$ naturally corresponds to a function $k_x \in \mathcal{H}$. We call k_x the *kernel function* of x, or the kernel function centered at x.

Theorem 3.1.2 (Riesz Representation Theorem)**.** *If ϕ is a bounded linear functional on a Hilbert space $\mathcal{H}$, then there is a unique $g \in \mathcal{H}$ such that*

$$\phi(x) = \langle g, f \rangle$$

for all $f \in \mathcal{H}$.

Corollary 1. *Let $\mathcal{H}$ be a RKHS on X. For every $x \in X$, there exists a unique $k_x \in \mathcal{H}$ such that*

$$\langle k_x, f \rangle = f(x)$$

for all $f \in \mathcal{H}$.

The kernel function of x is "reproducing" in the sense that its inner product with a function f reproduces the value of f at x.

Definition 3.1.3 (Reproducing Kernel)**.** The function $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined by

$$K(x, y) = k_y(x)$$

is called the *reproducing kernel* of $\mathcal{H}$.

The kernel K is symmetric, as the inner product is symmetric:

$$K(x, y) = k_y(x) = \langle k_y, k_x \rangle = \langle k_x, k_y \rangle = k_x(y) = K(y, x)$$

If we were working in a complex vector space, the kernel would have conjugate symmetry.

Theorem 3.1.4 (Equivalence Between Kernels and RKHS)**.** *Every RKHS has a unique reproducing kernel, and every reproducing kernel induces a unique RKHS.*

Proof. We have already seen by means of the Riesz Representation Theorem that every RKHS induces a unique kernel. The converse is a consequence of the Cauchy-Schwartz inequality, which states $\langle x, y \rangle \leq \|x\| \, \|y\|$. If K is a reproducing kernel on a Hilbert space $\mathcal{H}$, then

$$e_x(f) = \langle k_x, f \rangle \leq \|k_x\| \, \|f\| = \sqrt{K(x, x)} \cdot \|f\|$$

so e_x is bounded, and $\mathcal{H}$ is an RKHS. □

The existence of a reproducing kernel is sometimes called the *reproducing kernel property.*

We note that although our original definition of an RKHS involved its evaluation functionals, it turns out to be much easier to think about such a space in terms of its kernel function than its evaluation functionals.

3.1.1 Examples

We now look at some concrete examples of Reproducing Kernel Hilbert Spaces, building up from simple spaces to more complex ones.

Example: Linear Functions in $\mathbb{R}^d$ We begin with the simplest of all Reproducing Kernel Hilbert Spaces, Euclidean spaces. Consider $\mathcal{H} = \mathbb{R}^d$ with the canonical basis vectors $e_1, \ldots, e_d$ and the standard inner product:

$$\langle x, w \rangle = \sum_{i=1}^{n} x_i w_i$$

With the notation above, X is the discrete set $\{1, \ldots, d\}$, and $e_i \in \mathcal{H}$ is the kernel function

$$\langle e_i, x \rangle = x(i) = x_i$$

The reproducing kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is simply the identity matrix

$$K(i, j) = \langle e_i, e_j \rangle = \mathbf{1} i == j$$

so that for any $x, x' \in \mathbb{R}^d$, we have

$$K(x, x') = \langle x, x' \rangle$$

In general, for any discrete set X, the Hilbert space $L^2(X) = \{f \in \mathbb{R}^X : \sum_x |f(x)|^2 < \infty\}$ of square-summable functions has a RKHS structure induced by the orthonormal basis vectors $e_y(x) = \mathbf{1}\{x = y\}$.

Example: Feature Maps in $\mathbb{R}^p$ We can extend the previous example by considering a set of linearly independent maps $D = \{\phi_i\}_{i=1}^p$ for $\phi_i : X \to \mathbb{R}$. Let $\mathcal{H}$ be the span:

$$\mathcal{H} = \text{span}\ \{D\} = \{f : X \to \mathbb{R} : f(x) = \sum_{i=1}^{p} w_i \phi_i(x) \text{ for some } w \in \mathbb{R}^p\}$$

The maps ϕ_i are called *feature maps* in the machine learning community.

We define the inner product on $\mathcal{H}$ by

$$\langle x, x' \rangle_{\mathcal{H}} = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^p} = \sum_{i=1}^{p} \phi_i(x)\phi_i(x')$$

and the kernel $K : X \times X \to \mathbb{R}$ is simply

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^p}$$

Linear functions correspond to the case where $X = \{1, \ldots, d\}$, $p = d$, and $\phi_i(x) = x_i$.

Example: Polynomials One of the most common examples of feature maps are the polynomials of degree at most s in $\mathbb{R}^d$. For example, for $s = 2$ and $d = 2$,

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$$

with corresponding polynomial kernel

$$\begin{aligned} K(x, x') &= 1 + 2x_1x_1' + 2x_2x_2' + 2x_1x_2x_1'x_2' + x_1^2x_1'^2 + x_2^2x_2'^2 \\ &= (1 + \langle x, x' \rangle)^2 \end{aligned}$$

In general, the RKHS of polynomials of degree at most s in $\mathbb{R}^d$ has kernel $(1 + \langle x, x' \rangle)^s$ and is a space of degree $\binom{s+d}{d}$.

Example: Paley-Wiener spaces The Paley-Wiener spaces are a classical example of a RKHS with a *translation invariant* kernel, which is to say a kernel of the form $K(x, x') = K'(\|x - x'\|)$ for some function K'. Paley-Wiener spaces are ubiquitous in signal processing, where translation invariance is a highly desirable property.

Since we are interested in translation-invariance, it is natural to work in frequency space. Recall the Fourier transform:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i x\xi}\, dt$$

Consider functions with limited frequencies, which is to say those whose Fourier transforms are supported on a compact region $[-A, A]$. Define the Paley-Wiener space PW_A as

$$PW_A = \{\hat{f} : f \in L^2([-A, A])\}$$

where L^2 refers to square-integrable functions.

We can endow PW_A with the structure of an RKHS by showing that it is isomorphic (as a Hilbert space) to $L^2([-A, A])$. By the definition of PW_A, for every $\hat{f} \in PW_A$, there exists an $f \in L^2[(-A, A)]$ such that

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi}\, dx = \int_{-A}^{A} f(x) e^{-2\pi i x \xi}\, dx$$

We claim that this transformation, viewed as a map $L^2([-A, A]) \to PW_A$, is an isomorphism. It is clearly linear, so we need to show that it is bijective.

To show bijectivity, note that the functions $\{x \mapsto e^{2\pi i n x/A}\}_{n \in \mathbb{N}}$ form a basis for $L^2([-A, A])$. Then if $\hat{f}(n/A) = 0$ for every $n \in \mathbb{N}$, we have $f = 0$ almost everywhere, and vice-versa. Therefore $L^2([-A, A])$ and PW_A are isomorphic.

We can now give PW_A the inner product

$$\langle \hat{f}_1, \hat{f}_2 \rangle_{PW_A} = \langle f_1, f_2 \rangle_{L_2} = \int_{-A}^{A} f_1(x) f_2(x)\, dx$$

Since for any $\hat{f} \in PW_A$,

$$|\hat{f}(x)| = \left| \langle f, e^{2\pi i x \xi} \rangle_{L^2} \right| \leq \left\| e^{2\pi i x \xi} \right\|_{L^2} \|f\|_{L^2} = \sqrt{2A} \left\| \hat{f} \right\|$$

so the evaluation functionals $f \mapsto f(x)$ are bounded, and PW_A is an RKHS.

To obtain the kernel, we can use the fact that

$$\langle \hat{f}, \widehat{k_y} \rangle_{L^2} = \langle f, k_y \rangle_{PW_A} = f(y) = \langle \hat{f}, e^{2\pi i y t} \rangle_{L^2}$$

which gives by the inverse Fourier transform that $k_y(x) = \widehat{e^{2\pi i y \xi}}(x)$. Computing this integral gives the kernel:

$$\begin{aligned} K(x, y) = k_y(x) &= \int_{-A}^{A} e^{2\pi i (x-y)\xi}\, d\xi \\ &= \begin{cases} 2A & x = y \\ \sin(2\pi A(x-y))/(\pi(x-y)) & x \neq y \end{cases} \end{aligned}$$

This kernel is a transformation of the sinc function, defined as:

$$\operatorname{sinc}(x) = \begin{cases} 1 & x = 0 \\ \sin(x)/x & x \neq 0 \end{cases}, \qquad K(x, y) = 2A\operatorname{sinc}(2A\pi(x - y))$$

Example: Sobolev Spaces Sobolev spaces are spaces of absolutely continuous functions that arise throughout real and complex analysis.

A function $f : [0, 1] \to \mathbb{R}$ is *absolutely continuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that, if a finite sequence of pairwise disjoint sub-intervals $\{(x_k, y_k)\} \subset [0, 1]$ satisfies $\sum_k y_k - x_k < \varepsilon$, then $\sum_k |f(y_k) - f(x_k)| < \delta$.

Intuitively, absolutely continuous functions are those that satisfy the fundamental theorem of calculus. Indeed, the fundamental theorem of Lebesgue integral calculus states that the following are equivalent:

1. f is absolutely continuous
2. f has a derivative almost everywhere and $f(x) = f(a) + \int_a^x f'(t)dt$ for all $x \in [a, b]$.

Let $\mathcal{H}$ be the set of absolutely continuous functions with square-integrable derivatives that are 0 at 0 and 1:

$$\mathcal{H} = \{f : f' \in L^2([0, 1]),\ f(0) = f(1) = 0,\ f \text{ absolutely continuous}\}$$

We endow $\mathcal{H}$ with the inner product

$$\langle f, g\rangle = \int_0^1 f'(x)g'(x)dx$$

We see that the values of functions in $\mathcal{H}$ are bounded

$$\begin{aligned} |f(x)| &= \int_0^x f'(t)\,dt = \int_0^1 f'(t)\mathbf{1}\{t < x\}\,dt \\ &\leq \left(\int_0^1 f'(t)^2\,dt\right)^{1/2} \left(\int_0^1 \mathbf{1}\{t < x\}\,dt\right)^{1/2} = \|f\|\,\sqrt{x} \end{aligned}$$

so the evaluation functionals are bounded. It is simple to show that with this inner product, the space $\mathcal{H}$ is complete, so $\mathcal{H}$ is an RKHS.

We now compute the kernel k_x in a manner that is non-rigorous, but could be made rigorous with additional formalisms. We begin by integrating by parts:

$$\begin{aligned} f(x) = \langle f, k_x \rangle &= \int_0^1 f'(t) k_x'(t)\, dt = f(t)k_x'(t)|_0^1 - \int_0^1 f(t) k_x''(t)\, dt \\ &= -f(t)k_x''(t)\, dt \end{aligned}$$

We see that if k_x were to satisfy

$$-k_x''(t) = \delta_x(t), \quad k_x(0) = 0, \quad k_x(1) = 0$$

where δ_x is the Dirac delta function, it would be a reproducing kernel. Such a function is called the Green's function, and it gives us the solution:

$$k_x(t) = K(t,x) = \begin{cases} (1-x)t & t \le x \\ (1-t)x & x \le t \end{cases}$$

It is now easy to verify that

$$\begin{aligned} \langle f, k_x \rangle &= \int_0^1 f'(t) k_x'(t)\, dt \\ &= \int_0^x f'(t)(1-x)\, dt + \int_x^1 f'(t)(-x)\, dt \\ &= f(x) \end{aligned}$$

An Example from Stochastic Calculus In the above example, we considered a function f on $[0,1]$ with a square-integrable derivative f' and fixed the value of f to 0 and $t = 0, 1$. We found that the kernel $K(x,t)$ is given by $x(1-t)$ for $x < t$.

If the reader is familiar with stochastic calculus, this description might sound familiar. In particular, it resembles the definition of a Brownian bridge. This is a stochastic process X_t whose distribution equals that of Brownian motion conditional on $X_0 = X_1 = 0$. Its covariance function is given by $\text{Cov}(X_s, X_t) = s(1-t)$ for $s < t$.

Now consider the space $\mathcal{H}$ of functions for which we only require $f(0) = 0$:

$$\mathcal{H} = \{f : f' \in L^2([0,1]),\ f(0) = 0,\ f \text{ absolutely continuous}\}$$

If the previous example resembled a Brownian bridge, this example resembles Brownian motion. Indeed, by a similar procedure to the example above, one can show that the kernel function of $\mathcal{H}$ is given by

$$K(x,t) = \min(s,t)$$

which matches the covariance $\mathrm{Cov}(B_s, B_t) = \min(s,t)$ of Brownian motion.

This remarkable connection is no coincidence. Given a stochastic process X_t with covariance function R, it is possible to define a Hilbert space $\mathcal{H}$ generated by this X_t. A fundamental theorem due to Loeve [60] states that this Hilbert space is congruent to the Reproducing Kernel Hilbert space with kernel R.

Example: The Sobolev Space H^1 Consider the space

$$\mathcal{H} = H^1 = \{f : f \in L^2(\mathbb{R}), f' \in L^2(\mathbb{R}), f \text{ absolutely continuous}\}$$

endowed with the inner product

$$\langle f, g \rangle = \frac{1}{2} \int_{-\infty}^{\infty} f(t)g(t) + f'(t)g'(t)\, dt$$

which induces the norm

$$\|f\|_{H^1}^2 = \frac{1}{2}\left(\|f\|_{\mathcal{L}^2}^2 + \|f'\|_{\mathcal{L}^2}^2\right)$$

The resulting RKHS H^1, another example of a Sobolev space, may be understood in a number of ways.

From the perspective of the Paley-Wiener spaces example, it is a translation-invariant kernel best viewed in Fourier space. One can use Fourier transforms to show that $K(x,y) = \kappa(|x-y|)$, where $\hat{\kappa}(\xi) = \frac{2}{1+\xi}$. Then an inverse Fourier transform shows K is given by

$$K(x,y) = \frac{1}{2} e^{-|x-y|}$$

From the perspective of stochastic calculus, this space corresponds to the Ornstein–Uhlenbeck process

$$dX_t = -\theta\, X_t\, dt + \sigma\, dB_t$$

which is square-continuous but not square-integrable. The kernel function of $\mathcal{H}$ corresponds to the covariance function of the OU process:[1]

$$K(s,t) \propto \mathrm{Cov}(B_s, B_t) = \frac{\sigma^2}{2\theta} e^{-\theta|s-t|}$$

Finally, we note that we can generalize this example. For any $\gamma > 0$, the kernel

$$K(x,y) = \frac{1}{2} e^{-\gamma|x-y|}$$

is called the *exponential kernel*, and corresponds to the norm

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\gamma} \left(\|f\|_{\mathcal{L}^2}^2 + \|f'\|_{\mathcal{L}^2}^2 \right)$$

3.1.2 Structure

Thus far, we have defined an RKHS as a Hilbert space with the reproducing property and given a number of examples of such spaces. However, it is not yet clear *why* we need the reproducing property. Indeed, all of the examples above could have been presented simply as Hilbert spaces with inner products, rather than as RKHSs with kernels.

The best way of conveying the importance of the reproducing property would be to give an example of a Hilbert space that is not an RKHS and show that it is badly behaved. However, explicitly constructing such an example is impossible. It is equivalent to giving an example of an unbounded linear functional, which can only be done (non-constructively) using the Axiom of Choice.

One commonly and incorrectly cited example of a Hilbert space that is not an RKHS is $L^2(\Omega)$, the space of square-integrable functions on a domain Ω. This example is not valid because L^2 is technically not a set of functions, but rather a set of equivalence classes of functions that differ on sets of measure 0. Whereas L^2 spaces are not concerned with the values of functions on individual points (only on sets of positive measure), Reproducing Kernel Hilbert Spaces are very much concerned with the values of functions on individual points.[2] In this sense, RKHSs behave quite differently from L^2 spaces.

[1] Technically, an OU process with an initial condition drawn from a stationary distribution, or equivalently the limit of an OU process away from a strict boundary condition.

[2] The reader is encouraged to go back and check that all of the examples above (particularly Paley-Wiener spaces) are defined in terms of functions that are well-defined pointwise, rather than equivalence classes of functions.

Anti-Example This example illustrates the idea that the norm in L^2 does not control the function pointwise. Consider a sequence $f_n \in L^2([0,1])$ defined by

$$f_n(x) = \begin{cases} 1 & \frac{1}{2} - \frac{1}{n} \le x \le \frac{1}{2} + \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

As $n \to \infty$, it converges in L^2 norm to the 0 function. However, its value at 1/2 is always $f(1/2) = 1$. This is to say, there exist functions with arbitrarily small norm and unbounded values at individual points.

The purpose of the reproducing property of an RKHS is to prevent this type of behavior.

Theorem 3.1.5. *Let $\mathcal{H}$ be an RKHS on X. If $\lim_{n\to\infty} \|f_n - f\| = 0$, then $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in X$.*

Proof. By the existence of reproducing kernels and Cauchy-Schwartz,

$$|f_n(x) - f(x)| = |(f_n - f)(x)| = |\langle f_n - f, k_x \rangle| \le \|f_n - f\| \, \|k_x\|$$

so $\lim_{n\to\infty} |f_n(x) - f(x)| = 0$. □

We may also express K pointwise in terms of the basis of the underlying Hilbert space.

Theorem 3.1.6. *Denote by $\{e_s\}_{s\in S}$ a basis for the RKHS $\mathcal{H}$. Then*

$$K(x,y) = \sum_{s\in S} e_s(x) e_s(y)$$

where convergence is pointwise.

Proof. By the reproducing property,

$$k_y = \sum_{s\in S} \langle k_y, e_s \rangle e_s = \sum_{s\in S} e_s(y) e_s$$

where the sum converges in norm, and so converges pointwise. Then

$$K(x,y) = k_y(x) = \sum_{s\in S} e_s(y) e_s(x)$$

□

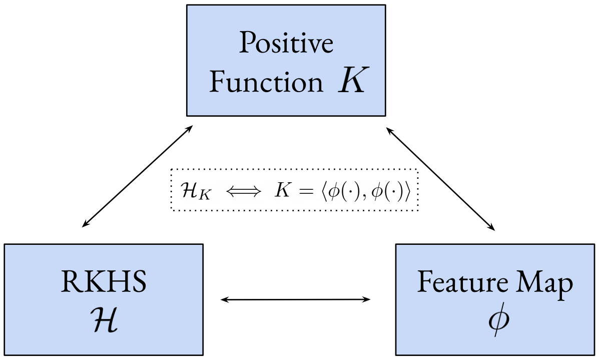

Positive Function K
$\mathcal{H}_K \iff K = \langle \phi(\cdot), \phi(\cdot) \rangle$
RKHS $\mathcal{H}$
Feature Map ϕ

Figure 3.2.1: An illustration of the equivalence between kernels, positive functions, and inner products of feature maps.

3.2 Kernels, Positive Functions, and Feature Maps

At this point, we are ready to fully characterize the set of kernel functions.

Definition 3.2.1 (Positive Function)**.** Let X be an arbitrary set. A symmetric function $K : X \times X \to$ is a *positive* function if for any n points $\{x_1, \ldots, x_n\}$ in X, the matrix $(K)_{ij} = K(x_i, x_j)$ is positive semidefinite. Equivalently, for any $c_1, \ldots, c_n$ in $\mathbb{R}$, we have

$$\sum_{i=1}^{n} \sum_{i=1}^{n} c_i c_j K(x_i, x_j)$$

Note: Positive functions are sometimes also called positive definite, positive semidefinite, nonnegative, or semipositive. We will use the term *positive* to mean ≥ 0, and the term *strictly positive* to mean > 0.

We now prove that there is a one-to-one correspondence between kernels and positive functions.

Theorem 3.2.2. *If $K = \langle \cdot, \cdot \rangle$ is the kernel of an RKHS $\mathcal{H}$, it is a positive function.*

Proof. First note that K is symmetric, as the inner product on $\mathcal{H}$ is symmetric. Second, we compute

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \langle \sum_{i=1}^{n} c_i x_i, \sum_{i=1}^{n} c_i x_i \rangle = \left\| \sum_{i=1}^{n} c_i x_i \right\|^2 \geq 0$$

□

The reverse direction is a celebrated theorem attributed to Moore.

Theorem 3.2.3 (Moore-Aronszajn Theorem)**.** *Let X be a set and suppose $K : X \times X \to \mathbb{R}$ is a positive function. Then there is a unique Hilbert space $\mathcal{H}$ of functions on $X \to \mathbb{R}$ for which K is a reproducing kernel.*

Proof. Define k_y by $k_y(x) = K(x, y)$. Note that if K were the kernel of an RKHS $\mathcal{H}$, then the span of the set $\{k_y\}_{y \in X}$ would be dense in $\mathcal{H}$, because if $\langle k_y, f \rangle = 0$ for all $y \in X$, then $f(y) = 0$ for all $x \in X$.

With this motivation, define V to be the vector space spanned by $\{k_y\}_{y \in X}$. Define the bilinear form $\langle \cdot, \cdot \rangle$ on V by

$$\langle \sum_i c_i k_{y_i}, \sum_i c_i' k_{y_i} \rangle = \sum_{i,j} c_i c_j' K(y_i, y_j)$$

We aim to show that $\langle \cdot, \cdot \rangle$ is an inner product. It is positive-definite, bilinear, and symmetric by the properties of K, so it remains to be shown that it is well defined. To do so, we need to check $f = 0 \iff \langle f, g \rangle = 0$ for all $g \in V$.

($\Longrightarrow$) If $\langle f, g \rangle = 0$ for all $g \in V$, letting $g = k_y$ we see $\langle f, g \rangle = f(y) = 0$ for all $y \in X$. Therefore $f = 0$.

($\Longleftarrow$) If $f = 0$, $\langle f, k_y \rangle = \sum_i c_i K(x_i, y) = f(y) = 0$. Since the k_y span V, each $g \in V$ may be expressed as a linear combination of the k_y, and $\langle f, g \rangle = 0$ for all $g \in V$.

Therefore $\langle \cdot, \cdot \rangle$ is well-defined and is an inner product on V. Moreover, we may produce the completion $\mathcal{G}$ of V by considering Cauchy sequences with respect to the norm induced by this inner product. Note that $\mathcal{G}$ is a Hilbert space.

All that remains is to identify a bijection between $\mathcal{G}$ and the set of functions $X \to \mathbb{R}$. Note that this is where an L^2 space fails to be an RKHS. Let $\mathcal{H}$ be the set of functions of the form $\overline{f}(x) = \langle f, k_x \rangle$, such that

$$\mathcal{H} = \{\overline{f} : f \in \mathcal{G}\}$$

and observe that elements of $\mathcal{H}$ are functions $X \to \mathbb{R}$. We see that if $\overline{f} = 0$, then $\langle f, k_x \rangle = 0$ for all $x \in X$, so $h = 0$. Therefore the mapping $f \mapsto \overline{f}$ is linear (by the properties of the inner product) and one-to-one. Thus, the space $\mathcal{H}$ with the inner product $\langle \overline{f}, \overline{g} \rangle_{\mathcal{H}} = \langle f, g \rangle_{\mathcal{G}}$ is a Hilbert space with the reproducing kernels $\overline{k_x}$ for $x \in X$. This is our desired RKHS. □

There is one final piece in the RKHS puzzle, the concept of feature spaces.

Let X be a set. Given a Hilbert space $\mathbb{F}$, not necessarily composed of functions $X \to \mathbb{R}$, a *feature map* is a function $\phi : X \to \mathbb{F}$. In machine learning, X and $\mathbb{F}$ are usually called the *data space* and the *feature space*, respectively. Above, we saw this example in the case $\mathcal{H} = \mathbb{R}^p$. Now ϕ may take values in an infinite-dimensional Hilbert space, but the idea remains exactly the same.

Given a feature map ϕ, we construct the kernel given by the inner product

$$K(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$$

or equivalently $\phi(x) = k_x$. As shown above, this kernel defines an RKHS on X.

Conversely, every kernel K may be written as an inner product $\langle \phi(\cdot), \phi(\cdot) \rangle$ for some feature map ϕ. In other words, the following diagram commutes:

$$\begin{tikzcd}
X \times X \arrow[d, "\phi"'] \arrow[dr, "K"] & \\
\mathbb{F} \times \mathbb{F} \arrow[r, "\langle \cdot, \cdot \rangle"] & \mathbb{R}
\end{tikzcd}$$

We note that the Hilbert space $\mathbb{F}$ and feature map ϕ above are not unique. However, the resulting Reproducing Kernel Hilbert Space, composed of functions $X \to \mathbb{R}$, is unique. In other words, although a feature map specifies a unique RKHS, a single RKHS may have possible feature map representations.

Theorem 3.2.4. *A function $K : X \times X \to \mathbb{R}$ is positive if and only if it may be written as $\langle \phi(\cdot), \phi(\cdot) \rangle$ for some Hilbert space $\mathbb{F}$ and some map $\phi : X \to \mathbb{F}$.*

Proof. We give a proof for finite-dimensional Hilbert spaces. It may be extended to the infinite-dimensional case with spectral operator theory, but we will not give all the details here.

First, suppose $K = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathbb{F}}$. Then for $v \in \mathbb{F}$,

$$\langle v, Kv \rangle = \sum_{i=1}^{n} v_i \sum_{i=1}^{n} \langle \phi(x_i), \phi(x_j) \rangle v_j = \langle \sum_{i=1}^{n} v_i \phi(x_i), \sum_{i=1}^{n} v_i \phi(x_i) \rangle \geq 0$$

so K is positive definite.

Second, suppose K is positive. Decompose it into $K = V \Lambda V^T$ by the spectral theorem, and let $\phi(x) = \Lambda^{1/2} V^T \mathbf{1}_x$. Then we have

$$\langle \phi(x), \phi(x') \rangle_{\mathbb{F}} = \langle \mathbf{1}_x, \mathbf{1}_{x'} \rangle_K = K(x, x')$$

so $K = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathbb{F}}$. □

We now have a full picture of the relationship between Reproducing Kernel Hilbert Spaces, positive-definite functions, and feature maps.

3.2.1 Geometry

One way to think of an infinite-dimensional RKHS is as a map $x \mapsto k_x$ that sends every point in X to a point $k_x : X \to \mathbb{R}$ in an infinite-dimensional feature space.

The kernel function K defines the geometry of the infinite-dimensional feature space.

Example: Gaussian Kernel Let $X = \mathbb{R}^d$ and consider the Gaussian kernel, perhaps the most widely used kernel in machine learning:

$$K(x, x') = e^{-\frac{1}{2}\|x - x'\|^2}$$

The kernel function k_x corresponding to a point x is a Gaussian centered at x. Due to its radial symmetry, this kernel is also called the *radial basis function* (RBF) kernel.

It turns out that explicitly constructing the RKHS for the Gaussian kernel is challenging (it was only given by [92] in 2006). However, since it is not difficult to show that K is a positive function, we can be sure that such an RKHS exists.

Let us look at its geometry. We see that each point $x \in X$ is mapped to a point k_x with unit length, as $\|k_x\|^2 = K(x,x) = 1$. The distance between two points k_x, k_y is:

$$\begin{aligned}\|k_x - k_y\|^2 &= K(x-y, x-y) = K(x,x) - 2K(x,y) + K(y,y) \\ &= 2\left(1 - e^{-\frac{1}{2}\|x-y\|^2}\right) < 2\end{aligned}$$

so any two points are no more than $\sqrt{2}$ apart.

Example: Min Kernel Consider the kernel $K(s,t) = \min(s,t)$ for $s, t \in \mathbb{R}$. This kernel induces a squared distance

$$\begin{aligned}d_K(s,t)^2 &= K(s,s) - 2K(s,t) + K(t,t) \\ &= s + t - 2\min(s,t) \\ &= \max(s,t) - \min(s,t) \\ &= |s-t|\end{aligned}$$

the square root of the standard squared Euclidean distance on $\mathbb{R}$.

In general, so long as the map $x \mapsto k_x$ is unique, the function

$$d_K(x,y) = \sqrt{K(x-y, x-y)} = \sqrt{K(x,x) - 2K(x,y) + K(y,y)}$$

is a valid distance metric on $\mathcal{H}$. In this sense, the kernel defines the similarity between two points x and y. From a feature map perspective, the distance is

$$d_K(x,y) = \|\phi(x) - \phi(y)\|$$

This metric enables us to understand the geometry of spaces that, like the RKHS for the Gaussian Kernel, are difficult to write down explicitly.

3.2.2 Integral Operators

We now take a brief detour to discuss the relationship between kernels and integral operators. This connection will prove useful in Chapter 5.

We say that a kernel $K : X \times X \to \mathbb{R}$ is a *Mercer kernel* if it is continuous and integrable.[3] That is, $K \in L^2(X \times X)$, meaning

[3] The notation used throughout the literature is not consistent. It is common to see “Mercer kernel” used interchangeably with “kernel”. In practice, nearly every kernel of interest is a Mercer kernel.

$\int_X \int_X K(x,x')\,dx\,dx' < \infty$.

Suppose that X is compact and define the integral operator $I_K : L^2(X) \to L^2(X)$ by

$$I_K f(x) = \int_X K(x,x') f(x')\,dx'$$

It is not difficult to show that I_K is linear, continuous, compact, self-adjoint, and positive. Linearity follows from the linearity of integrals, continuity from Cauchy-Schwartz, compactness from an application of the Arzelà–Ascoli theorem, self-adjointness from an application of Fubini's theorem, and positivity from the fact that the integral fI_kf is a limit of finite sums of the form $\sum_{i,j} f(x_i)K(x_i,x_j)f(x_j) \geq 0$.

Since I_K is a compact, positive operator, the spectral theorem states that there exists a basis of $L^2(X)$ composed of eigenfunctions of I_K. Denote these eigenfunctions and their corresponding eigenvalues by $\{\phi_i\}_{i=1}^\infty$ and $\{\lambda_i\}_{i=1}^\infty$, respectively. Mercer's theorem states that one can decompose K in this basis:

Theorem 3.2.5 (Mercer)**.**

$$K(x,y) = \int_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$$

where the convergence is absolute and uniform over $X \times X$.

This theorem is not challenging to prove, but it requires building significant machinery that would not be of further use. We direct the interested reader to [73] (Section 98) for a detailed proof.

3.3 Tikhonov Regularization and the Representer Theorem

Having built our mathematical toolkit, we return now to machine learning. Our goal is to minimize the regularized empirical risk $\hat{\mathcal{E}}(f(x),y) + \lambda R(f,x,y)$ over a space $\mathcal{H}$.

Let $\mathcal{H}$ be an RKHS, as we are concerned with the values of functions pointwise. Let R be the norm $\|f\|_K^2 = K(f,f)$, as its purpose is to measure the complexity of a function.

Denote our data by $S = \{(x_i,y_i)\}_{i=1}^N$, and let $\hat{\mathcal{E}}(f,x,y)$ be the sum of a loss function $L(f(x_i), y_i)$ over the data. Our learning problem is then

$$\arg\min_{f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i) + \lambda \|f\|_K^2 \tag{3.1}$$

where $\lambda > 0$. This general framework is known as *Tikhonov regularization.*

The Representer Theorem reduces this infinite-dimensional optimization problem to a finite-dimensional one. It states that our desired solution is a linear combination of the kernel functions on the data points.

Theorem 3.3.1 (Representer Theorem)**.** *Let $\mathcal{H}$ be an RKHS on a set X with kernel K. Fix a set of points $S = \{x_1, x_2, \dots, x_N\} \subset X$. Let*

$$J(f) = L(f(x_1), \dots, f(x_n)) + R(\|f\|_{\mathcal{H}}^2)$$

and consider the optimization problem

$$\min_{f\in\mathcal{H}} J(f)$$

where R is nondecreasing. Then if a minimizer exists, there is a minimizer of the form

$$f = \sum_{i=1}^{N} \alpha_i k_{x_i}$$

where $\alpha_i \in \mathbb{R}$. Moreover, if P is strictly increasing, every minimizer has this form.

Proof. The proof is a simple orthogonality argument.

Consider the subspace $T \subset \mathcal{H}$ spanned by the kernels at the data points:

$$T = \text{span}\ \{k_{x_i} : x_i \in S\}$$

Since S is a finite dimensional subspace, so it is closed, and every $f \in \mathcal{H}$ may be uniquely decomposed as $f = f_T + f_\perp$, where $f_T \in T$ and $f_\perp \in T^\perp$.

By the reproducing property, we may write $f(x_i)$ as

$$\begin{aligned} f(x_i) &= \langle f, k_{x_i}\rangle = \langle f_T, k_{x_i}\rangle + \langle f_\perp, k_{x_i}\rangle = \langle f_T, k_{x_i}\rangle \\ &= f_T(x_i) \end{aligned}$$

Also note

$$R(\|f\|^2) = R(\|f_T\|^2 + \|f_\perp\|^2) \geq R(\|f_T\|^2)$$

Then $J(f)$ may be written as

$$\begin{aligned} J(f) &= L(f(x_1), \dots, f(x_n)) + R(\|f\|^2) = L(f_T(x_1), \dots, f_T(x_n)) + R(\|f\|^2) \\ &\geq L(f_T(x_1), \dots, f_T(x_n)) + R(\|f_T\|^2) \\ &= J(f_T) \end{aligned}$$

Therefore, if f is a minimizer of J, f_T is also a minimizer of J, and f_T has the desired form. Furthermore, if R is strictly increasing, the $\geq$ above may be replaced with $>$, so f cannot be a minimizer of J unless $f = f_T$. □

If L is a convex function, then a minimizer to Equation 3.1 exists, so by the Representer Theorem it has the form

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)$$

Practically, it converts the learning problem from one of dimension d (that of the RKHS) to one of dimension N (the size of our data set). In particular, it enables us to learn even when d is infinite.

3.4 Algorithms

With the learning problem now fully specified, we are ready to look at algorithms.

Regularized Least Squares Regression

In regularized least squares regression, we aim to learn a function $f : X \to \mathbb{R}$ minimizing the empirical risk with the loss function $L(f(x), y) = (f(x) - y)^2$. In other words, the learning problem is:

$$\arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2$$

where $(x_i, y_i) \in X \times \mathbb{R}$ are our (training) data.

By the Representer Theorem, the solution f of this learning problem may be written:

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)$$

We now solve for the parameters $\alpha = (\alpha_1, \dots, \alpha_N)$.

For ease of notation, we write $x = (x_1, \dots, x_N)$, $y = (y_1, \dots, y_N)$. Denote by K the $N \times N$ kernel matrix on the data, also called the *Gram matrix*: $K = (K_{ij})(K(x_i, x_j))$. With this notation, we have

$$(f(x_1), \dots, f(x_n)) = K\alpha \qquad \text{and} \qquad \|f\|_K^2 = \alpha^T K \alpha$$

so our objective may be written as

$$\arg\min_{f \in \mathcal{H}} \frac{1}{N}(K\alpha - y)^T (K\alpha - y) + \lambda \alpha^T K \alpha \tag{3.2}$$

To optimize, we differentiate with respect to α, set the result to 0, and solve:

$$0 = \frac{2}{N} K(K\alpha^* - y) + 2\lambda K \alpha^* = K((K + \lambda N I)\alpha^* - y) \tag{3.3}$$

Since K is positive semidefinite, $(K + \lambda N I)$ is invertible, and

$$\alpha^* = (K + \lambda N I)^{-1} y$$

is a solution. Therefore

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)$$

with $\alpha = (K + \lambda N I)^{-1} y$ is a minimizer of the learning problem.

If $X = \mathbb{R}^d$ with the canonical inner product, the Gram matrix is simply $K = XX^T$, where X is the $N \times d$ matrix of data. Then α^* becomes $\alpha^* = (XX^T + \lambda N I)^{-1} y$ and the minimizer of the learning problem may be written as

$$X^T (XX^T + \lambda N I)^{-1} y \tag{3.4}$$

A Woodbury matrix identity states that for any matrices U, V of the correct size, $(I + UV)^{-1} = I - U(I + VU)^{-1} V$. The expression above may then be written as

$$(X^T X + \lambda N I)^{-1} X^T y \tag{3.5}$$

which is the familiar solution to a least squares linear regression.

Comparing Equations 3.4 and 3.5, we see that the former involves inverting a matrix of size $N \times N$, whereas the latter involves inverting a matrix of size $d \times d$. As a result, if $d > N$, it may be advantageous to use 3.4 even for a linear kernel.

A Note on Uniqueness: The process above showed that $\alpha^* = (K + \lambda N I)^{-1} y$ is a solution to Equation 3.2, but not that it is unique. Indeed, if the rank of K is less than N, multiple optimal $\alpha \in \mathbb{R}^d$ may exist. However, the function $f \in \mathcal{H}$ constructed from these α will be the same. To see this, note that Equation 3.3 shows that for any optimal α, we have $\alpha = (K + \lambda N I)^{-1} - y + \delta$, where $K\delta = 0$. Therefore for any two optimal α, α' we have

$$\left\| f - f' \right\|^2 = (\alpha - \alpha')^T K (\alpha - \alpha') = 0$$

and so $f = f'$.

Regularized Logistic Regression

Regularized logistic regression, which is a binary classification problem, corresponds to the logistic loss function

$$\log(1 + e^{-y_i f(x_i)})$$

where the binary labels y are represented as $\{-1, 1\}$. Our objective is then

$$\arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i f(x_i)}) + \lambda \left\| f \right\|_K^2$$

Our solution takes the form given by the Representer Theorem, so we need to solve

$$\arg\min_{\alpha \in \mathbb{R}^N} \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i (K\alpha)_i}) + \lambda \alpha^T K \alpha$$

for α. Unfortunately, unlike for least squares regression, this equation has no closed form. Fortunately, it is convex, differentiable, and highly amenable to gradient-based optimization techniques (e.g. gradient descent). These optimization methods are not a focus of this thesis, so we will not go into further detail, but we note that they are computationally efficient and widely used in practice.

Regularized Support Vector Machines

Regularized support vector classification, also a binary classification problem, corresponds to the hinge loss function

$$L_{sup}(f(x), y) = \max(0, 1 - yf(x)) = (1 - yf(x))_+$$

where $y_i \in \{-1, 1\}$. As always, our objective is

$$\arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i f(x_i)}) + \lambda \|f\|_K^2$$

and our solution takes the form given by the Representer Theorem. Like with logistic regression, we solve

$$\arg\min_{\alpha \in \mathbb{R}^N} \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i (K\alpha)_i}) + \lambda \alpha^T K \alpha$$

for α by computational methods. The one caveat here is that we need to use "subgradient-based" optimization techniques rather than gradient-based techniques, as the gradient of the hinge loss is undefined at 0.

The Kernel Trick

Suppose we have an algorithm $\mathcal{A}$ where the data x_i are only used in the form $\langle x_i, \cdot \rangle$. In this case, we can *kernelize* the algorithm by replacing its inner product with a kernel K. This process, known as the *kernel trick*, effectively enables us to work in infinite-dimensional feature spaces using only finite computational resources (i.e. only computing the kernel functions K).

3.4.1 Building Kernels

In practice, applying kernel methods translates to building kernels that are appropriate for one's specific data and task. Using task-specific kernels, it is possible to encode one's domain knowledge or inductive biases into a learning algorithm. The problem of automatically selecting or building a kernel for a given task is an active area of research known as *automatic*

Name	Periodic	Kernel	Areas of Application
Linear	✗	$x^T x'$	Ubiquitous
Polynomial	✗	$(c + x^T x')^p$	Ubiquitous
Gaussian	✓	$e^{-\frac{1}{2\sigma}\lVert x-y\rVert^2}$	Ubiquitous
Exponential	✓	$e^{-\sigma\lVert x-y\rVert}$	Ubiquitous
Tanh	✗	$\tanh(\sigma x^T x' + b)$	Neural networks
Dirichlet	✓	$\frac{\sin((n+1/2)(x-x'))}{2\pi\sin((x-x')/2)}$	Fourier analysis
Poisson	✓	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x-x')+1}$	Laplace equation in 2D
Sinc	✓	$\frac{\sin(\sigma(x-x'))}{(x-x')}$	Signal processing
Rational Quadratic	✓	$\sigma^2\left(1+\frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha}$	Gaussian processes
Exp-Sine-Squared	✓	$\sigma^2\exp\left(-\frac{2\sin^2(\pi\lvert x-x'\rvert/p)}{\ell^2}\right)$	Gaussian processes
Matérn Kernel	✓	$\sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\lvert x-x'\rvert}{\rho}\right)^{\nu}K_\nu\left(\sqrt{2\nu}\frac{\lvert x-x'\rvert}{\rho}\right)$	Gaussian processes

Table 3.4.1: Examples of commonly used kernel functions.

kernel selection.

Although building kernels for specific tasks is outside the scope of this thesis, we give below a few building blocks for kernel construction. Using these building blocks, one can create complex kernels from simpler ones.

Properties Let K, K' be kernels on X, and let f be a function on X. Then the following are all kernels:

- $K(x,x') + K'(x,x')$
- $K(x,x') \cdot K'(x,x')$
- $f(x)K(x,x')f(x')$
- $K(f(x), f(x'))$
- $\exp(K(x,x'))$
- $\frac{K(x,x')}{\sqrt{K(x,x)}\sqrt{K(x',x')}}$, called the *normalized* version of K

We remark that all these properties may be thought of as properties of positive functions.

Kernels from Probability Theory A few interesting kernels arise from probability theory. For events A, B, the following are kernels:

- $K(A, B) = P(A \cap B)$ is a kernel.
- $K(A, B) = P(A \cap B) - P(A)P(B)$ is a kernel.
- $H(X) + H(X') - H(X, X')$

At first glance, the mutual information $I(X, X')$ also looks like a kernel. The question of whether it was a kernel was solved in 2012 by Jakobsen [46], who showed that $I(X, X')$ is a kernel if and only if $\dim(X) \leq 3$.

Common Kernels in Machine Learning Common examples of kernels are given in Table 3.4.1. Still more examples of kernels are available at this link.

4
Graphs and Manifolds

We now turn our attention from the topic of Reproducing Kernel Hilbert Spaces to an entirely new topic: the geometry of graphs and Riemannian manifolds. The next and final chapter will combine these two topics to tackle regularized learning problems on graphs and manifolds.

The purpose of this chapter is to elucidate the connection between graphs and manifolds. At first glance, these two mathematical objects may not seem so similar. We usually think about graphs in terms of their combinatorial properties, whereas we usually think about manifolds in terms of their topological and geometric properties.

Looking a little deeper, however, there is a deep relationship between the two objects. We shall see this relationship manifest in the Laplacian operator, which emerges as a natural operator on both graphs and manifolds. The same spectral properties of the Laplacian enable us to understand the combinatorics of graphs and the geometry of manifolds.

This chapter explores how the two Laplacians encode the structures of their respective objects and how they relate to one another. By the end of the chapter, I hope the reader feels that graphs are discrete versions of manifolds and manifolds are continuous versions of graphs.

Related Work & Outline

Numerous well-written references exist for spectral graph theory [22, 79] and for analysis on manifolds [18], but these topics are usually treated independent from one another.[1] One notable exception is [14], illustratively titled "How is a graph like a manifold?". This paper examines a different aspect of the graph-manifold connection from the one examined here; whereas [14] is concerned with group actions on complex manifolds and their connections to graph combinatorics, this chapter is concerned with spectral properties of the Laplacian on both manifolds and graphs.

Rather than discuss graphs and then manifolds, or vice-versa, we discuss the two topics with a unifying view. Throughout, we highlight the relationship between the Laplacian spectrum and the concept of connectivity of a graph or manifold.

We assume that the reader is familiar with some introductory differential geometry (i.e. the definition of a manifold), but has not necessarily seen the Laplacian operator on either graphs or manifolds before.

4.1 Smoothness and the Laplacian

As seen throughout the past two chapters, we are interested in finding smooth functions. On a graph or a manifold, what does it mean to be a smooth function? The Laplacian holds the key to our answer.

Let $G = (V, E)$ be a connected, undirected graph with edges E and vertices V. The edges of the graph can be weighted or unweighted (with nonnegative weights); we will assume it is *unweighted* except where otherwise specified. When discussing weighted graphs, we denote by w_{ij} the weight on the edge between nodes i and j.

A real-valued function on G is a map $f : V \to \mathbb{R}$ defined on the vertices of the graph. Note that these functions are synonymous with vectors, as they are of finite length.

Intuitively, a function on a graph is smooth if its value at a node is similar to its value at each of the node's neighbors. Using squared

[1] The literature on Laplacian-based analysis of manifolds is slightly more sparse the spectral graph theory literature. For the interested reader, I highly recommend [18].

difference to measure this, we arrive at the following expression:

$$\sum_{(i,j)\in E} (f(i) - f(j))^2 \tag{4.1}$$

This expression is a symmetric quadratic form, so there exists a symmetric matrix $\mathbf{L}$ such that

$$\mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{(i,j)\in E} (f(i) - f(j))^2$$

where $\mathbf{f} = (x(1), \ldots, x(n))$ for $n = |V|$.

We call $\mathbf{L}$ the Laplacian of the graph G. We may think of $\mathbf{L}$ as a functional on the graph that quantifies the smoothness of functions.

The Laplacian of a weighted graph is defined similarly, by means of the following quadratic form:

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \sum_{(i,j)\in E} w_{ij}(x(i) - x(j))^2$$

Notation: Some texts work with the normalized Laplacian $\mathcal{L}$ rather than the standard Laplacian $\mathbf{L}$. The normalized Laplacian is given by $D^{-1/2}\mathbf{L}D^{-1/2}$, where D is the diagonal matrix of degrees of vertices (i.e. $D_{ii} = \deg(i)$).

We now turn our attention to manifolds. Let $(\mathcal{M}, g)$ be a Riemannian manifold of dimension n. As a refresher, this means that $\mathcal{M}$ is a smooth manifold and g is a map that smoothly assigns to each $x \in \mathcal{M}$ an inner product $\langle \cdot, \cdot \rangle_{g_x}$ on the tangent space $T_x\mathcal{M}$ at x. For ease of notation, when it is clear we will write $\mathcal{M}$ in place of $(\mathcal{M}, g)$ and $g_x(\cdot, \cdot)$ in place of $\langle \cdot, \cdot \rangle_{g(x)}$.

Suppose we wish to quantify the smoothness of a function $f : \mathcal{M} \to \mathbb{R}$ at a point $x \in \mathcal{M}$. A natural way of doing this would be to look at the squared norm $\|\nabla f\|^2$ of the gradient of f at x. This quantity is analogous to the squared difference between a node's value and the values of its neighbors in the graph case. Informally, if we write $\|\nabla f\|^2$ as $f\nabla \cdot \nabla f$, it looks like a quadratic form. As in the graph case, we associate this form with an operator Δ.

Formally, we define Δ as the negative divergence of the gradient, written as $\Delta = -\nabla \cdot \nabla$ or $-\operatorname{div} \nabla$ or $-\nabla^2$. We call Δ the Laplacian or Laplace-Beltrami operator on the manifold $\mathcal{M}$.

Notation: Some texts define Δ as $\operatorname{div} \nabla$, without a negative sign. In these texts, the Laplace-Beltrami operator is negative semidefinite and its eigenvalue equation is written as $\Delta f = -\lambda f$ rather than $\Delta f = \lambda f$. Here, we adopt the negated version for simplicity and for consistency with the graph literature, where the Laplacian is positive semidefinite.

Since $\|\nabla f(x)\|^2$ describes the smoothness of a function f at x, integrating it over the entire manifold gives a notion of the smoothness of f on $\mathcal{M}$:

$$\int_{\mathcal{M}} \|\nabla f(x)\|^2 \, dx$$

This quantity (technically 1/2 of this quantity) is called the *Dirichlet energy* and denoted by $E[f]$. It plays a role analogous to Equation 4.1 on the graph, and occurs throughout physics as a measure of the variability of a function. In fact, the Laplace operator may be thought of as the functional derivative of the Dirichlet energy.

4.1.1 More Definitions and Properties

Readers familiar with graph theory or analysis may have noticed that the definitions given above are not the most common ways to introduce Laplacians on either graphs or manifolds.

Usually, one defines the Laplacian of a graph G in terms of the adjacency matrix A.[2] The Laplacian is given by

$$\mathbf{L} = D - A$$

where $D_{ii} = \deg(i)$ is the diagonal matrix of degrees of nodes. The normalized laplacian is then:

$$\mathcal{L} = I - D^{-1/2} A D^{-1/2}$$

A simple computation shows that these definition and our original one are

[2] At first glance, the adjacency matrix might seem to be the most natural matrix to associate to a graph. However, for a variety of reasons, the Laplacian in general turns out to be much more connected to the fundamental combinatorial properties of the graph. The one notable exception to this rule is in studying random walks, where the powers and spectrum of the adjacency matrix define the behavior and equilibrium state of the random walk.

equivalent:

$$\begin{aligned} x^T(D-A)x &= x^TDx + x^TAx \\ &= \sum_{i=1}^{n} \deg(i)x_i^2 - \sum_{(i,j)\in E} 2x_ix_j \\ &= \sum_{i=1}^{n} \sum_{(i,j)\in E} x_i^2 - \sum_{(i,j)\in E} 2x_ix_j \\ &= \sum_{(i,j)\in E} (x_i^2 + x_j^2 - 2x_ix_j) \\ &= \sum_{(i,j)\in E} (x_i - x_j)^2 \\ &= x^T\mathbf{L}x \end{aligned}$$

Some basic properties of the Laplacian, although not obvious from the definition $\mathbf{L} = D - A$, are obvious given the quadratic form definition. Namely, $\mathbf{L}$ is symmetric and positive semi-definite, since for any x,

$$x^T\mathbf{L}x = \sum_{(i,j)\in E} (x_i - x_j)^2 \geq 0$$

As a result, all eigenvalues of $\mathbf{L}$ are non-negative. We can also see that the smallest eigenvalue is 0, corresponding to an eigenfunction that is a (non-zero) constant function.

Turning to manifolds, the Laplacian Δ is also usually introduced in a different manner from the one above. In the context of multivariable calculus, it is often defined as:

$$\Delta f = -\frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial y^2} - \frac{\partial^2 f}{\partial z^2}$$

which is easily verified to be equal to $\operatorname{div} \nabla f$ in $\mathbb{R}^N$. This coordinate-wise definition can be extended to the local coordinates of a Riemannian manifold with metric tensor g:

$$\Delta = -\frac{1}{\sqrt{|\det g|}} \sum_{i,j=1}^{n} \left(g^{ij} \sqrt{|\det g|} \frac{\partial}{\partial x_j} \right) \tag{4.2}$$

However, if one would like to work with coordinates on a manifold, it is much more natural to work in the *canonical* local coordinates. To switch to these coordinates, we use the exponential map $\exp_p : T_p\mathcal{M}(=\mathbb{R}^n) \to \mathcal{M}$, which is a local diffeomorphism between a neighborhood of a point $p \in \mathcal{M}$ and a neighborhood of 0 in the tangent space $T_p\mathcal{M}$. This coordinate map gives a canonical identification of a neighborhood of p with $\mathbb{R}^N$, called geodesic normal coordinates. In geodesic normal coordinates, $g_{ij} = \delta_{ij}$ and $\frac{\partial g_{ij}}{\partial x_k} = 0$, so the formula for Δ resembles the formula in Euclidean space.

Finally, we should note that yet another way to define the Laplacian Δ is as the trace of the Hessian operator H:

$$\Delta = \mathrm{Tr}(H)$$

where the Hessian H at p is $\nabla_p(df)$, the gradient of the differential of f. Note that since the Hessian is coordinate-free (i.e. invariant under isometries), this relation shows us that Laplacian is coordinate-free.

4.1.2 Examples

Below, we present a few examples of Riemannian manifolds and graphs along with their Laplacians.

Example: $\mathbb{R}^n$ The most ordinary of all Riemannian manifolds is $\mathbb{R}^n$ with the Euclidean metric $g = \langle \cdot, \cdot \rangle_{\mathbb{R}^n}$. In matrix form, g is the identity matrix of dimension n: $g_{ij} = \delta_{ij}$ and $\det g = 1$. Following formula 4.2, we have

$$\Delta_{g,\,\mathbb{R}^n} = -\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$$

which is the familiar form of the divergence of the gradient in $\mathbb{R}^n$.

Example: S^1 The simplest nontrivial Riemannian manifold is the circle $S^1 \subset \mathbb{R}^2$ with the metric induced by $\mathbb{R}^2$. We may parameterize the circle as $(\cos(\theta), \sin(\theta))$, with the resulting metric $g = d\theta^2$ (induced from $\mathbb{R}^2$ as $dx^2 + dy^2 = dr^2 + r^2\, d\theta^2 = d\theta^2$). In matrix form, g is simply the 1-dimensional matrix (1). Consequently,

$$\Delta_{g,\,S^1} = -\frac{\partial^2}{\partial \theta^2}$$

as above. A similar result holds for all one-dimensional manifolds.

Example: Cycle Graph A simple graph similar to the smooth circle above is the cycle graph. The Laplacian $\mathbf{L}$ of a cycle graph G with n vertices is given by:

$$\mathbf{L} = D - A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & \ddots & 0 \\ 0 & \ddots & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

Readers familiar with numerical analysis might note that this matrix resembles the (negated) second-order discrete difference operator

$$\frac{\partial^2 u}{\partial x^2} \approx -\frac{-u_{i+1} + 2u_i - u_{i-1}}{\delta x}$$

which suggests a connection to the manifolds above. As we will see later, the Laplacian spectra of the circle and the cycle graph are closely related.

Example: S^2 Consider the 2-sphere parameterized in spherical coordinates with the metric induced from $\mathbb{R}^3$:

$$T : [0, \pi) \times [0, 2\pi) \to S^2$$

$$T(\theta, \phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$$

Changing to spherical coordinates shows that the metric is given by

$$g = dx^2 + dy^2 + dz^2 = (dx^2 + dy^2) + dz^2 = d\theta^2 + \sin^2\theta d\phi$$

so in matrix form g is

$$g(\theta, \phi) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}$$

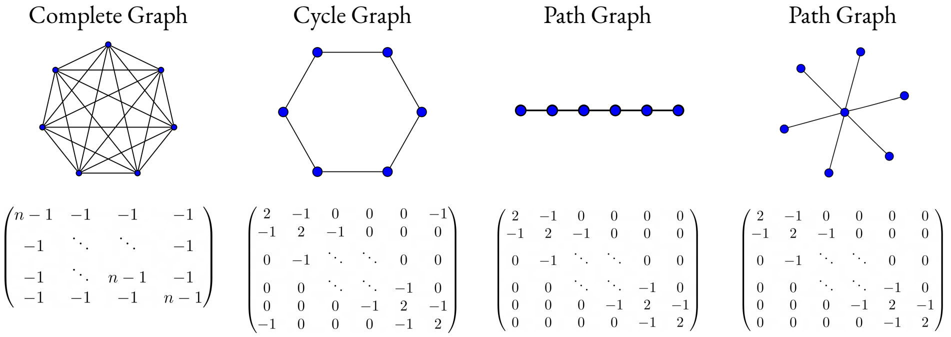

Complete Graph
Cycle Graph
Path Graph
Path Graph
$\begin{pmatrix} n-1 & -1 & -1 & -1 \\ -1 & \ddots & \ddots & -1 \\ -1 & \ddots & n-1 & -1 \\ -1 & -1 & -1 & n-1 \end{pmatrix}$
$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}$
$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}$
$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}$

Figure 4.1.1: A few classic graphs and their Laplacians.

with determinant $\det g = \sin^2\theta$. Then by formula 4.2, the Laplacian is

$$\begin{aligned}\Delta &= -\frac{1}{\sqrt{\det g}}\left(\frac{\partial}{\partial\theta}\left(g_{\theta\theta}\sqrt{\det g}\frac{\partial}{\partial\theta}\right)+\frac{\partial}{\partial\phi}\left(g_{\phi\phi}\sqrt{\det g}\frac{\partial}{\partial\phi}\right)\right)\\ &= -\frac{1}{\sin\theta}\frac{}{\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)-\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\end{aligned}$$

This expression enables us to work with the eigenvalue equation $\Delta f = \lambda f$ in spherical coordinates, a useful tool in electrodynamics and thermodynamics.

Example: Three More Fundamental Graphs   Figure 4.1.1 shows examples of three fundamental graphs—the complete graph, star graph, and path graph—along with their Laplacians.

Example: Flat Torus   An n-dimensional torus is a classic example of a compact Riemannian manifold with genus one, which is to say a single "hole".

Topologically, a torus $\mathbb{T}$ is the product of spheres, $S^1 \times \cdots \times S^1 = (S^1)^n$. Equivalently, a torus may be identified with $\mathbb{R}^n/\Gamma$, where Γ is an n-dimensional lattice in $\mathbb{R}^n$ (a discrete subgroup of $\mathbb{R}^n$ isomorphic to $\mathbb{Z}^n$). [3] That is to say, we can identify the torus with a (skewed and stretched) square in $\mathbb{R}^2$ conforming to specific boundary conditions (namely, that

[3] Concretely, Γ is the set of linear combinations with integer coefficients of a basis $\{e_1, e_2, \ldots, e_n\}$ of $\mathbb{R}^n$.

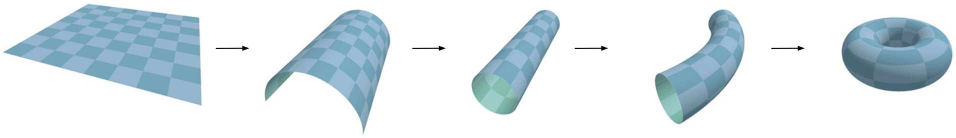


Figure 4.1.2: A fun illustration of how a torus may be created from a square in the plane with periodic boundary conditions.

opposite sides of the square are the same). We call the torus with $\Gamma = \mathbb{Z}^n$ the standard torus.

When endowed with the product metric from S^1 (i.e. the n-times product of the canonical metric on S^1), a torus is called the flat torus.[4] As the Laplacian is locally defined by the metric, the Laplacian of any flat surface is the same as the Laplacian in Euclidean space, restricted to functions that are well-defined on the surface.

Intuitively, the flat metric makes the torus look locally like $\mathbb{R}^n$. Among other things, this means that angles and distances work as one would expect in $\mathbb{R}^n$; for example, the interior angles of a triangle on a flat torus add up to π degrees.

Example: Torus Embedded in $\mathbb{R}^3$ The flat metric is not the only metric one can place on a torus. On the contrary, it is natural to picture a torus embedded in $\mathbb{R}^3$, with the familiar shape of a donut (Figure 4.1.2). The torus endowed with the metric induced from $\mathbb{R}^3$ is a different Riemannian manifold from the flat torus.

The torus $\mathbb{T}$ embedded in R^3 with minor radius (i.e. the radius of tube) r and outer radius (i.e. the radius from center of hole to center of tube) $R > r$ may be parameterized as

$$T : [0, 2\pi) \times [0, 2\pi) \to \mathbb{T}^2$$

$$T(\theta, \phi) = ((R + r\cos\theta)\cos\phi, (R + r\cos\theta)\sin\phi, r\sin\theta)$$

[4] In general, a manifold is said to be flat if it has zero curvature at all points. Examples of other spaces commonly endowed with a flat metric include the cylinder, the Möbius band, and the Klein bottle.

The metric g inherited from $\mathbb{R}^3$ is

$$\begin{aligned}
g &= dx^2 + dy^2 + dz^2 \\
&= d((R + r\cos\theta)\cos\phi)^2 + d((R + r\cos\theta)\sin\phi)^2 + d(r\sin\theta)^2 \\
&= (d\phi\sin\phi(-(r\cos\theta + R)) - r\cos\phi d\theta\sin\theta)^2 \\
&\quad + (d\phi\cos\phi(r\cos\theta + R) - r\sin\phi d\theta\sin\theta)^2 + r^2 d\theta^2\cos^2\theta \\
&= (R + r\cos\theta)^2 d\phi^2 + r^2 d\theta^2
\end{aligned}$$

and so the corresponding matrix (g_{ij}) is

$$g(\theta, \phi) = \begin{pmatrix} r^2 & 1 \\ 1 & (R + r\cos\theta)^2 \end{pmatrix}$$

The Laplacian of the torus embedded in $\mathbb{R}^3$ is then

$$\Delta f = -\frac{1}{\sqrt{|\det g|}} \sum_{i,j=1}^{n} \left(g^{ij} \sqrt{|\det g|} \frac{\partial}{\partial x_j} \right) \tag{4.3}$$

$$= -r^{-2} (R + r\cos\theta)^{-1} \frac{\partial}{\partial\theta} (R + r\cos\theta) \frac{\partial}{\partial\theta} f - (R + r\cos\theta)^{-2} \frac{\partial^2}{\partial\phi^2} f \tag{4.4}$$

Whereas the distances and angles on the flat torus act similarly to those in $\mathbb{R}^2$, distances and angles on the embedded torus act as we would expect from a donut shape in $\mathbb{R}^3$. For example, the sum of angles of an triangle drawn on a flat torus is always π, but this is not true on the torus embedded in $\mathbb{R}^3$.[5]

More formally, the embedded torus is diffeomorphic to the flat torus but not isomorphic to it: there exists a smooth and smoothly invertible map between them, but no such map that preserves distances. In fact, there does not exist a smooth embedding of the flat torus in $\mathbb{R}^3$ that preserves its metric. [6]

[5] A triangle drawn on the "inside" of the torus embedded in $\mathbb{R}^3$ has a sum of angles that is less than π, whereas a triangle drawn on the "outside" has a sum of angles that is greater than π. Although we will not discuss Gaussian curvature in this text, we note that this sum of angles is governed by the curvature of the surface, which is negative on the inside of the torus and positive on the outside. As another example, the sum of angles of a triangle on the 2-sphere, which has positive Gaussian curvature, is $\frac{3\pi}{2}$.

[6] For the interested reader, we remark that it is known that there does not even exist a smooth metric-preserving (i.e. isometric) C^2 embedding of the flat torus in R^3. However, results of Nash from 1950 show that there does exist an isometric C^1 embedding. In 2012, the first explicit construction of such an embedding was found; its structure resembles that

4.2 Lessons from Physics

We would be remiss if we introduced the Laplacian without discussing its connections to physics. These connections are most clear for the Laplacian on manifolds, which figures in a number of partial differential equations, including the ubiquitous heat equation.

Example: Fluid Flow (Manifolds) Suppose we are physicists studying the movement of a fluid over a continuous domain D. We model the fluid as a vector field v. Experimentally, we find that the fluid is incompressible, so $\operatorname{div} v = 0$, and conservative, so $v = -\nabla u$ for some function u (the potential). The potential then must satisfy

$$\Delta u = 0$$

This is known as Laplace's Equation, and its solutions are called harmonic functions.

Example: Fluid Flow (Graphs) Now suppose we are modeling the flow of a fluid through pipes that connect a set of reservoirs. These reservoirs and pipes are nodes and edges in a graph G, and we may represent the pressure at each reservoir as a function u on the vertices.

Physically, the amount of fluid that flows through a pipe is proportional to the difference in pressure between its vertices, $u_i - u_j$. Since the total flow into each vertex equals the total flow out, the sum of the flows along a vertex i is 0:

$$0 = \sum_{j \in N(i)} u_i - u_j \tag{4.5}$$

Expanding this gives:

$$\begin{aligned} 0 &= \sum_{j \in N(i)} u_j - \sum_{j \in N(i)} u_i = \deg(i) u_i - \sum_{j \in N(i)} u_j \\ &= ((D - A)u)_i = (\mathbf{L}u)_i \end{aligned}$$

We find that $\mathbf{L}u$ is 0, a discrete analogue to the Laplace equation $\Delta u = 0$.

Equivalently, Equation 4.5 means that each neighbor is the average of its

of a fractal [15].

neighbors:

$$u_i = \frac{1}{\deg(i)} \sum_{j \in N(i)} u_j$$

We can extend this result from 1-hop neighbors to k-hop neighbors, by noting that each of the 1-hop neighbors is an average of their own neighbors and using induction.

While this result is obvious in the discrete case, it is quite non-obvious in the continuous case. There, the analogous statement is that a harmonic functions equals its average over a ball.

Theorem 4.2.1 (Mean Value Property of Hamonic Functions)**.** *Let $u \in C^2(\Omega)$ be a harmonic function on an open set Ω. Then for every ball $B_r(x) \subset \Omega$, we have*

$$u(x) = \frac{1}{|B_r(x)|} \int_{B_r(x)} u(x)\, dx = \frac{1}{|\partial B_r(x)|} \int_{\partial B_r(x)} u(x)\, dx$$

where ∂B_r denotes the boundary of B_r.

If one were were to only see this continuous result, it might seem somewhat remarkable, but in the context of graphs, it is much more intuitive.

For graphs, the converse of these results is also clear. If a function u on a graph is equal to the average of its k-hop neighbors for any k, then the sum in Equation 4.5 is zero, so $\mathbf{L}u = 0$ and u is harmonic. For manifolds, it is also true that if u equals its average over all balls centered at each point x, then u is harmonic.

Example: Gravity Written in differential form, Gauss's law for gravity says that the gravitational field g induced by an object with mass density ρ satisfies

$$\nabla g = -4\pi G \rho$$

where G is a constant. Like our model of a fluid above, the gravitational field is conservative, so $g = -\nabla \phi$ for some potential function ϕ. We then see

$$\Delta \phi = 4\pi G \rho$$

Generally, a partial differential equation of the form above

$$\Delta u = f$$

is known as the Poisson equation.

Note that if the mass density is a Dirac delta function, meaning that all the mass is concentrated at a single point, the solution to this expression turns out to be $\phi(r) = -Gm/r$, which is Newton's law of gravitation.

Example: Springs Consider a graph in which each node exerts upon its neighbors an attractive force. For example, we could imagine each vertex of the graph as a point a $2D$ plane connected to its neighbors by a spring.

Hooke's Law states that the potential energy of a spring is $\frac{k}{2}x^2$, where $x \in \mathbb{R}^2$ is the amount the spring is extended or compressed from its resting displacement. Working in the $2D$ plane, the length of the spring is the difference $\|\mathbf{x}_i - \mathbf{x}_j\|$ where $\mathbf{x}_i = (x_i, y_i)$ and $\mathbf{x}_j = (x_j, y_i) \in \mathbb{R}^2$ are the positions of the two nodes.

If the resting displacement of each spring is 0, the potential energy in the (i, j) spring is $\frac{k}{2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2$. The total potential energy of our system is sum of the energies in each spring:

$$\sum_{(i,j)\in E} \frac{k}{2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \propto x^T\mathbf{L}x + y^T\mathbf{L}y$$

We see that finding a minimum-energy arrangement corresponds to minimizing a Laplacian quadratic form. If we were working in $\mathbb{R}^1$ instead of $\mathbb{R}^2$, the expression above would coincide exactly with our traditional notion of the Laplacian $x^T\mathbf{L}x$.

Harmonic Functions As seen repeatedly above, we are interested in harmonic functions, those for which $\mathbf{L} = 0$. However, on a finite graph, all such functions are constant!

We can see this from our physical system of springs with resting displacement 0. Intuitively, if G is connected, the springs will continue pulling the vertices together until they have all settled on a single point, corresponding to a constant function. Alternatively, if $x^T\mathbf{L}x = 0$, then each term $(x(i) - x(j))^2$ in the Laplacian quadratic form must be 0, so x must be constant on each neighborhood. Since G is connected, $x(i)$ must then be constant for all vertices i.

Nonetheless, all is not lost. Interesting functions emerge when we place additional conditions on some of the vertices of the graph. In the case of the spring network, for example, we can imagine nailing some of the

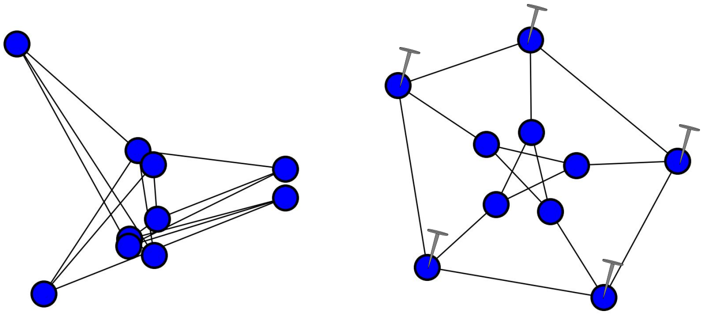


Figure 4.2.1: An illustration of Tutte's Theorem. On the left, we embed a graph into the plane by placing its vertices at random positions. On the right, we show the same graph embedded by taking one of its faces, nailing them in place, and letting the others settle into an arrangement with minimal potential energy.

vertices onto specific positions in the $2D$ plane. If we let this system come to equilibrium, the untethered vertices will settle into positions in the convex hull of the nailed-down vertices, as shown in Figure 4.2.1.

In fact, a famous theorem of Tutte [84] states that if one fixes the edges of a face in a (planar) graph and lets the others settle into a position that minimizes the total potential energy, the resulting embedding will have no intersecting edges.

Theorem 4.2.2 (Tutte's Theorem)**.** *Let $G = (V, E)$ be a 3-connected, planar graph. Let F be a set of vertices that forms a face of G. Fix an embedding $F \to \mathbb{R}^2$ such that the vertices of F form a strictly convex polygon. Then this embedding may be extended to an embedding $V \to \mathbb{R}^2$ of all of G such that*

1. *Every vertex in $V \setminus F$ lies at the average of its neighbors.*
2. *No edges intersect or self-intersect.*

The statements above all have continuous analogues. Like a harmonic function on a finite graph, a harmonic function on a compact manifold without boundary (a *closed* manifold) is constant.

Theorem 4.2.3. *If f is a harmonic function on a compact boundaryless region D, f is constant.*

On a region with boundary, a harmonic function is determined entirely by its values on the boundary.

Theorem 4.2.4 (Uniqueness of harmonic functions)**.** *Let f and g be harmonic functions on a compact region D with boundary ∂D. If $f = g$ on ∂D, then $f = g$ on D.*

As a result, if a harmonic function is zero on its boundary, it is zero everywhere. This result is often stated in the form of the maximum principle.

Theorem 4.2.5 (Maximum Principle)**.** *If f is harmonic on a bounded region, it attains its absolute minimum and maximum on the boundary.*

The maximum principle corresponds to the idea that if we nail the vertices of the face of a graph to the plane, the other nodes will settle inside of their convex hull; if every point is the average of its neighbors, the maximum must be attained on the boundary.

EXAMPLE: MORE FLUIDS Returning to continuous fluids, suppose we are interested in understanding how a fluid evolves over time. For example, we may be interested in the diffusion of heat over a domain D. This process is governed by the ubiquitous heat equation:

$$\partial_t u(x,t) = \Delta u(x,t)$$

One common approach to solving this equation is to guess a solution of the form $u(x,t) = \alpha(t)\phi(x)$ and proceed by separation of variables. This yields:

$$\frac{\Delta\phi(x)}{\phi(x)} = -\frac{\alpha'(t)}{\alpha(t)}$$

which implies that

$$\alpha' = -\lambda\alpha \qquad \text{and} \qquad \Delta\phi = \lambda\phi$$

for some $\lambda \in \mathbb{R}$. The equation on the left yields $\alpha(t) = Ce^{-\lambda t}$, and the equation on the right shows that λ is an eigenvalue of Δ. This second equation is called the Helmholtz equation, and it shows that the eigenvalues of the Laplacian enable us to understand the processes it governs. Note also that the Laplace equation is a special case of the Helmholtz equation with $\lambda = 0$.

We discuss the heat equation (on both manifolds and graphs) in more detail in section 4.5. Before doing so, we need to understand the eigenvalues and eigenvectors of the Laplacian operator.

4.3 The Laplacian Spectrum

Our primary method of understanding the Laplacian will be by means of its eigenvalues, or spectrum.

We denote the eigenvalues of the Laplacians $\mathbf{L}$ and Δ by λ_i, with $\lambda_1 \leq \lambda_2 \leq \cdots$. We use the same symbols for both operators, but will make clear at all times which operator's eigenvalues we are referring to. In the graph case these are finite ($\mathbf{L}$ has n eigenvalues counting multiplicities), whereas in the case of a manifold they are infinite.

We have seen that $\mathbf{L}$ and Δ are self-adjoint positive-definite operators, so their eigenvalues are non-negative. By the spectral theorem, the eigenfunctions are orthonormal and form a basis for the Hilbert Space of L^2 functions on their domain. For a manifold $\mathcal{M} \subset \mathbb{R}^n$, the eigenfunctions form a basis for $L^2(\mathcal{M})$, and for a graph $G = (V, E)$, they form a basis for $L^2(V)$ (i.e. bounded vectors in $\mathbb{R}^n$).

We have also already seen that the constant function $\mathbf{1}$ is an eigenfunction of the Laplacian corresponding to eigenvalue $\lambda_1 = 0$.

Notation: Unfortunately, graph theorists and geometers use different conventions for the eigenvalues. Graph theorists number the eigenvalues $\lambda_1, \lambda_2, \ldots$, with $\lambda_1 = 0$, and prove theorems about the "second eigenvalue" of the Laplacian. Geometers number the eigenvalues $0, \lambda_1, \ldots$, and prove theorems about the "first eigenvalue" of the Laplacian. We will use the convention from spectral graph theory throughout this text.

Can you hear the shape of a drum? A famous article published in 1966 in the American Mathematical Monthly by Mark Kac asked "Can you hear the shape of a drum?" [51] The sounds made by a drumhead correspond to their frequencies, which are in turn determined by the eigenvalues of the Laplacian on the drum (a compact planar domain). If the shape of the drum is known, the problem of finding its frequencies is the Helmholtz equation above. Kac asked the inverse question: if the eigenvalues of the Laplacian are known, is it always possible to reconstruct the shape of the underlying surface? Formally, if D is a compact manifold with boundary on the plane, do the solutions of $\Delta u + \lambda u = 0$ with the boundary condition $u|_{\partial D} = 0$ uniquely determine D?

The problem remained unsolved until the early 1990s, when Gordon, Webb and Wolpert answered it negatively [39]. The simple counterexample

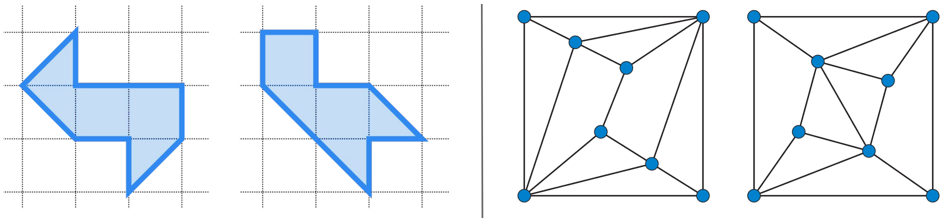


Figure 4.3.1: The two domains on the left have the same Laplacian spectrum, but are not isomorphic. The same is true of the two graphs on the right.

they presented is shown in Figure 4.3.1.

Nonetheless, the difficulty of proving this fact demonstrates just how much information the eigenvalues contain about the Laplacian. Indeed, Kac proved that the eigenvalues of Δ on a domain encode many geometric properties, including the domain's area, perimeter, and genus.

Similarly, it is not possible to reconstruct the structure of a graph from the eigenvalues of its Laplacian (Figure 4.3.1).[7]

4.3.1 Examples of Laplacian Spectra

Below, we give examples of the eigenvalues and eigenfunctions of a number of the manifolds and graphs from subsection 4.1.2.

Example: $\mathbb{C}^n$ and $\mathbb{R}^n$ In $\mathbb{C}^n$, the eigenvalue equation $\Delta f = \lambda f$ for the standard Laplacian $\Delta = -\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}$, is satisfied by the complex exponentials. In other words, the eigenfunctions of Δ are the functions $x \mapsto e^{i\sqrt{\lambda}x_i}$ for any $\lambda \geq 0$, where $\lambda = 0$ corresponds as usual to the constant function.

In $\mathbb{R}^n$, both the real and imaginary parts of the complex exponentials satisfy $-\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} f = \lambda f$. These are sine and cosine functions of the form $\sin(\sqrt{\lambda}x_i)$ and $\cos(\sqrt{\lambda}x_i)$, and as above every real λ in the continuous region $[0, \infty)$ is an eigenvalue.

[7] Also, if graphs with identical spectra were isomorphic, we would have a polynomial time solution to the graph isomorphism problem, the problem of determining whether two finite graphs are isomorphic. The graph isomorphism problem is neither known to be solvable in polynomial time nor known to be NP-complete.

EXAMPLE: S^1 The circle S^1, which inherits its metric from $\mathbb{R}^2$, looks locally like $\mathbb{R}^1$ but is globally periodic. The spectrum of its Laplacian are the functions on S^1 that solve

$$-\frac{\partial^2}{\partial\theta_i^2} f = \lambda f \tag{4.6}$$

which is to say they are the solutions to this equation in $\mathbb{R}^1$ that are also periodic with period 2π. These solutions take the form

$$f(\theta) = e^{ik\theta}$$

for $k \in \mathbb{Z}$. The real and imaginary parts of this expression yield the full set of eigenfunctions

$$f(\theta) = 1, \qquad f(\theta) = \sin(k\theta), \qquad f(\theta) = \cos(k\theta), \qquad \text{for } k = \{1, 2, \dots\}$$

with corresponding eigenvalues $0, k^2, k^2$ for $k \in \{1, 2, \dots\}$.

From another perspective, S^1 is locally like $\mathbb{R}^1$, so a sine/cosine wave with any wavelength locally satisfies Equation 4.6, but in order for it to be well-defined globally, its wavelength must be a multiple of 2π. Consequently, whereas the spectrum of Δ in $\mathbb{R}^1$ is continuous, the spectrum of Δ in S^1 is discrete. Consistent with this intuition, one can prove that all closed manifolds have discrete spectra, whereas non-compact manifolds may have continuous spectra.

Additionally, consider a circle with a non-unit radius r. From polar coordinates, we can see that the Riemannian metric is $g = r\,d\theta$ and the Laplacian becomes

$$\Delta f = -\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) - \frac{1}{r^2}\frac{\partial^2 f}{\partial\theta^2}$$

which has eigenvalues $0, k^2, k^2$ for $k \in \{1, 2, \dots\}$. As we increase the radius of our circle, we see that the spectrum becomes more dense in $\mathbb{R}$, and as it goes to infinity, we fill the entire region $[0, \infty)$, which is the spectrum of $\mathbb{R}^1$.

Example: Cycle Graph   As computed above, the Laplacian of the cycle graph is given by

$$\mathbf{L} = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

In Figure 4.3.2, we compute its eigenfunctions numerically for $n = 30$ and 100 vertices and plot the first six eigenfunctions. Comparing these to the plots of the eigenfunctions of the cycle graph, we see that the (scaled) eigenfunctions of the cycle graph approach those of the circle!

In this way, the cycle graph is a discrete version of a circle.

Example: Flat Torus   We saw previously that with the flat metric, the n-dimensional torus looks like a linearly transformed square in $\mathbb{R}^n$ with periodic boundary conditions. Formally, we have $\mathbb{T}^n = \mathbb{R}^n/\Gamma$ for an n-dimensional lattice Γ generated by a basis $\{e_1, \dots, e_n\}$ of $\mathbb{R}^n$.

To compute its eigenvalues, let Γ^* be the dual lattice, defined as $\{x \in \mathbb{R}^n : \langle x, y\rangle \in \mathbb{Z} \forall y \in \Gamma\}$. Just as with the other flat manifolds ($a\mathbb{R}^n$ and S^1) above, the solutions to eigenvalue equation $\Delta f = \lambda f$ solve $\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} f(x) = \lambda f$, so they are complex exponentials:

$$x \mapsto e^{2\pi i \langle x,y\rangle} \qquad \text{for all} \qquad y \in \Gamma$$

The real and imaginary parts yield the eigenfunctions $1, x \mapsto \sin(2\pi i\langle x, y\rangle)$, $x \mapsto \cos(2\pi i\langle x, y\rangle)$ for $y \in \Gamma^*$, which form a basis for $L^2(\mathbb{T}^n)$. The corresponding eigenvalues are $0, 4\pi^2|y|^2, 4\pi^2|y|^2$, similar to those on the circle S^1.

Example: Embedded Torus   We computed the Laplcaian of the 2-torus with the metric induced from $\mathbb{R}^3$, rather than the flat metric, in Equation 4.4. Its eigenvalue equation is then

$$\Delta f = -r^{-2} \left(R + r\cos\theta\right)^{-1} \frac{\partial}{\partial\theta} \left(R + r\cos\theta\right) \frac{\partial}{\partial\theta} f - \left(R + r\cos\theta\right)^{-2} \frac{\partial^2}{\partial\phi^2} f = \lambda f$$

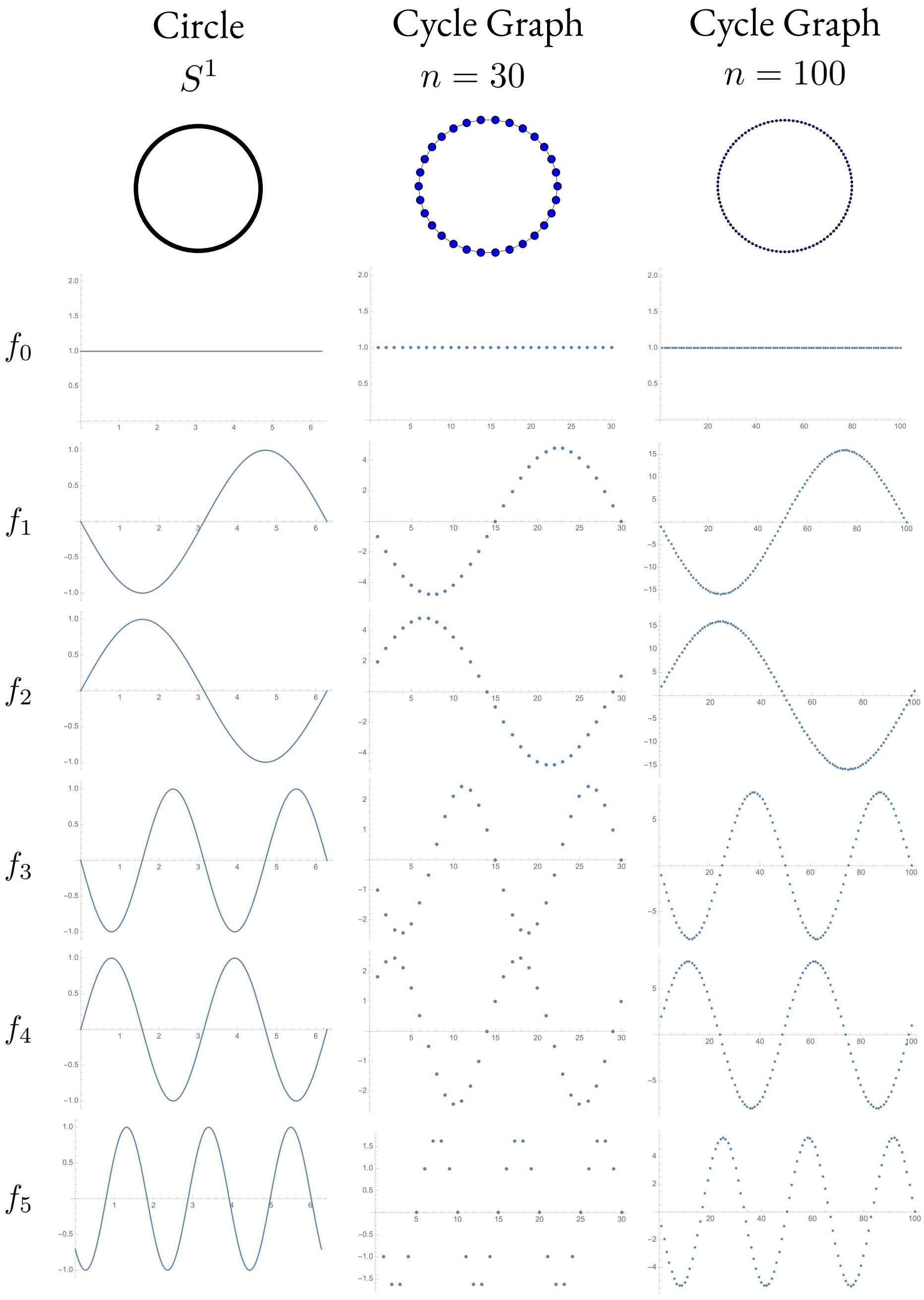


Figure 4.3.2: An illustration comparing the first six eigenfunctions of the circle and the cycle graph with $n = 30, 100$.

As this equation is separable, we consider a solution of the form $\psi(\theta, \phi) = a(\theta)e^{ik\phi}$ for $k \in \{1, 2, \dots\}$. Simplifying, we obtain

$$-\frac{1}{r^2}a''(\theta) + \frac{\sin\theta}{r + R\cos\theta}a'(\theta) + \frac{k^2}{(r + R\cos\theta)^2}a(\theta) = \lambda a(\theta)$$

which is an ordinary differential equation in a with periodic boundary conditions, solvable for given values of r and R. Note that each non-constant eigenvalue has multiplicity at least 2, corresponding to the real and imaginary parts of $e^{ik\phi}$, as with the flat torus and the circle.

Example: More Fundamental Graphs Recall from Figure 4.1.1 the Laplacians of the fully connected graph and the star graph on n vertices.

The eigenvalues of the complete graph, apart from $\lambda_1 = 0$, are n with multiplicity $n - 1$. As we shall see shortly, a graph's eigenvalues tell us about its connectedness, and the fully-connected graph has the largest eigenvalues.

The star graph has eigenvalues $\lambda_1 = 0$, $\lambda_n = n$, and $\lambda_i = 1$ for $1 < i < n$. Note that the second eigenvector, λ_2, is small. The graph is connected, but is "close" to being disconnected in the sense that if the middle vertex were removed, it would be entirely disconnected.

A star graph is an instance of a complete bipartite graph: its vertices can be divided into two subsets such that each vertex is connected (only) to the vertices of the other subset. In general, denoting by $K_{m,n}$ the complete bipartite graph with subsets of size m and $n - m$, the Laplacian $\mathbf{L}_{K_{m,n}}$ has eigenvalues 0, n, m, and $n + m$ with multiplicies $1, m - 1, n - 1$, and 1, respectively.

This result is a consequence of the following key lemma.

Lemma 4.3.1. *Let G be a simple graph. Let $\overline{G}$ be its complement, the graph on the same vertices as G such that each edge is included in $\overline{G}$ if and only if it is not in G. Denote the eigenvalues of the Laplacian $\boldsymbol{L}_G$ of G by $0 = \lambda_1 \leq \cdots \leq \lambda_n$. Then the eigenvalues of the Laplacian $\boldsymbol{L}_{\overline{G}}$ of $\overline{G}$ are*

$$0, n - \lambda_n, n - \lambda_{n-1}, \dots, n - \lambda_2$$

Proof. Let $v_1, \dots, v_n$ be orthonormal eigenvectors of $\mathbf{L}_G$ corresponding to

$\lambda_1, \dots, \lambda_n$. The sum of the Laplacians of G and $\overline{G}$ is

$$\begin{aligned}\mathbf{L}_G + \mathbf{L}_{\overline{G}} &= D_G - A_G + D_{\overline{G}} - A_{\overline{G}} = (D_G + D_{\overline{G}}) - (A_G + A_{\overline{G}}) \\ &= nI - J\end{aligned}$$

where J is the matrix of all $1s$. Now consider $\mathbf{L}_{\overline{G}} v_i$. If v_i is the constant vector, $\mathbf{L}_{\overline{G}} v_i = 0$. If it is not constant, it is orthogonal to the constant vector, so $Jv_i = 0$ and

$$\mathbf{L}_{\overline{G}} v_i = (nI - J - \mathbf{L}_G)v_i = nv_i - 0 - \lambda_i v_i = (n - \lambda_i)v_i$$

Therefore the eigenvalues of $\mathbf{L}_{\overline{G}}$ are $0, n - \lambda_n, n - \lambda_{n-1}, \dots, n - \lambda_2$. Also, its set of eigenvectors is the same as that of $\mathbf{L}_G$. □

From this lemma, it is quick to deduce the eigenvectors of the complete graph and $K_{m,n}$. The complete graph is the complement of the empty graph, which has eigenvalues $0^{(n)}$, so its eigenvalues are $0, n^{(n-1)}$. $K_{m,n}$ is the complement of the union of two complete graphs on n and m vertices. It is simple to show that the eigenvalues of the union of two graphs is the union of their eigenvalues, so the eigenvalues of the union are $0^{(2)}, n^{(n-1)}, m^{(m-1)}$. Then by the lemma the eigenvalues of $K_{m,n}$ are $0, n^{(m-1)}, m^{(n-1)}, n$.

Moreover, since the eigenvalues of every graph are nonnegative, the lemma shows that n is the largest that an eigenvalue of a graph with n vertices can be. In this way, the complete graph has the largest eigenvalues.

4.3.2 A Note on Boundaries

Before proceeding, we take a moment to address the concept of manifolds with boundary, as the reader likely has or will encounter such structures in the Riemannian geometry literature. We emphasize that finite graphs are analogous to *closed* (i.e. compact and boundaryless) manifolds, rather than those with boundary. A number of results in this text hold for manifolds with boundary and noncompact manifolds, but we make no guarantees.

For manifolds with boundary, the eigenfunctions of the Laplacian depends on both the underlying domain and the conditions placed on the boundary. For example, Kac's original "shape of a drum" question specified the boundary condition $u|_{\partial D} = 0$. This condition is the first of the two most widely-studied boundary conditions, *Dirichlet boundary* *conditions* and *Neumann boundary conditions*.[8]

Dirichlet boundary conditions require that the function be zero on its boundary:

$$\Delta u = \lambda u \text{ on } D, \qquad u|_{\partial D} = 0$$

Neumann boundary conditions require that the function's derivative be zero on its boundary:

$$\Delta u = \lambda u \text{ on } D, \qquad \frac{\partial u}{\partial \nu}|_{\partial D} = 0$$

where ν is the unit outward normal to ∂D.

To use the example of heat flow, Dirichlet boundary conditions correspond to a closed system in which no heat is allowed to enter or leave the system, whereas Neumann boundary conditions correspond to a system with a constant flow of heat at each point in the boundary.

These two types of boundary conditions only have graph analogues in the setting of *infinite* graphs. On finite graphs, fixing the value of a set of vertices determines a unique solution to $\Delta f = \lambda f$. Analogues of Dirichlet and Neumann boundary-value problems on infinite graphs is an active area of research [41, 47].

4.3.3 The Rayleigh Characterization of Eigenvalues

There are many ways of characterizing the eigenvalues of an operator. One particularly useful characterization is the Rayleigh quotient, which enables us to express eigenvalues as the solutions to optimization problems.

We begin in the setting of graphs. Let $\mathbf{A}$ be a self-adjoint matrix with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$. The Rayleigh quotient of a vector x is the expression

$$R(x) = \frac{x^T \mathbf{A} x}{x^T x}$$

where the denominator functions as a normalization factor. The Courant-Fischer Theorem states that λ_1 minimizes this expression over all nonzero x, λ_2 minimizes it over all x orthogonal to the first eigenvector, λ_2 maximizes it over x orthogonal to the first two eigenvectors, and so on.

[8] Although less common, other types of boundary conditions include Robin, Mixed, and Cauchy conditions. Each of these is different a combination of Dirichlet and Neumann boundary conditions (Robin is a linear combination, Mixed is a piecewise combination, and Cauchy imposes both at once).

Theorem 4.3.1 (Courant-Fischer). *The k-th smallest eigenvalue λ_k of the self-adjoint matrix $\boldsymbol{A}$ is given by*

$$\lambda_k = \min_{S \subset \mathbb{R}, \dim(S)=k} \max_{x \in S, x \neq 0} \frac{x^T \boldsymbol{A} x}{x^T x} \tag{4.7}$$

where S is a subspace of $\mathbb{R}^n$.

The proof is given in Appendix A.2.1.

For a Laplacian $\mathbf{L}$ of a graph G, the first eigenvalue $\lambda_1 = 0$ corresponds to the constant vector $\mathbf{1}$. We then immediately have what is known as the Rayleigh characterization of λ_2.

Corollary 2. *The first nonzero eigenvalue λ_2 of $\boldsymbol{L}$ is given by*

$$\lambda_2 = \min_{\|x\|^2=1, x \perp \boldsymbol{1}} x^T \boldsymbol{L} x$$

In what should not be an enormous surprise at this point, the Rayleigh quotient has an analogue on manifolds:

$$R(f) = \frac{\int_{\mathcal{M}} |\nabla f|^2 \, dV}{\int_{\mathcal{M}} f^2 \, dV} = \frac{\langle \nabla f, \nabla f \rangle}{\langle f, f \rangle}$$

where dV is the volume form on the manifold. The eigenvalues are given by the same optimization problem:

$$\lambda_1 = 0, \qquad \lambda_2 = \min \left\{ R(f) : \int_{\mathcal{M}} f \, dV = \langle f, \mathbf{1} \rangle = \int_{\mathcal{M}} f \, dV = 0 \right\}$$

The first eigenvalue is 0, corresponding to a constant eigenfunction, and the next largest eigenvalue is the minimizer of the Rayleigh quotient over all functions orthogonal to a constant function.[9] Subsequent eigenvalues $\lambda_3, \lambda_4, \dots$ of $\mathcal{M}$ may be obtained by a similar process as in the graph case.

$$\lambda_k = \min \left\{ R(f) : \langle f, f_i \rangle = 0 \quad \forall \quad i < k f_i \right\}$$

where f_i denotes the eigenfunction corresponding to the i-th eigenvalue λ_i.

[9] Technically, this minimization is taken over all functions f in the Sobolev space $H^1(\mathcal{M})$ corresponding to $\mathcal{M}$.

4.4 Eigenvalues and Connectivity

The Laplacian spectrum is closely related to the notion of connectedness.

4.4.1 The First Eigenvalues

The multiplicity of the first (zero) eigenvalue of the Laplacian gives the number of connected components of its corresponding graph or manifold.

Lemma 4.4.1. *The number of connected components of a graph G equals the multiplicity of the* 0 *eigenvalue of* $\boldsymbol{L}$*.*

Proof. First, suppose f is an eigenfunction of $\mathbf{L}$ corresponding to 0. Then $f\mathbf{L}f = \sum_{(i,j)\in E}(f(i) - f(j))^2 = 0$. In order for this sum to be 0, if f is nonzero on a vertex v, it must take the same value on every vertex connected to v. Then f must be constant on each component, meaning the multiplicity of the eigenvalue 0 is at most the number of connected components.

Second, note that for each connected component of the graph, the characteristic function of the component is an eigenfunction, so the multiplicity of the eigenvalue 0 is at least the number of connected components. □

For simplicity, we assume from now on that the graphs/manifolds we are discussing are connected, so λ_1 has multiplicity 1.

The second eigenvector λ_2 tells us about the connectivity of the graph or manifold in a different way from λ_1. Whereas λ_1 tells us whether the graph is connected at all, λ_2 gives us a sense of *how* connected the graph is. Informally, if λ_2 is small, then the graph is weakly connected, whereas if λ_2 is large, the graph is strongly connected. We have already seen one example of this idea above: a graph is fully connected if and only if λ_2 is as large as possible ($\lambda_2 = n$).

Graph theorists call λ_2 the *algebraic connectivity* of a graph. It is also sometimes referred to as *Fiedler value* for Czech mathematician Miroslav Fiedler, who was among the first to give bounds on λ_2.

Geometers call λ_2 the *fundamental tone* of a manifold. This name is derived from the fact that if we imagine a vibrating manifold , λ_2 is its leading frequency of oscillation.

4.4.2 Eigenvalue Bounds

We have seen that we can understand the structure of graphs and manifolds by looking at the eigenvalues of their Laplacians. In general, however, it is challenging to obtain analytic expressions for these eigenvalues.

Instead, most work is dedicated to proving and tightening bounds on these eigenvalues. The Rayleigh characterization of eigenvalues is useful because it gives us a simple method of obtaining an upper bound on λ_2: for any f, the Rayleigh quotient $\frac{\langle f, \mathbf{L}f \rangle}{\langle f, f \rangle}$ bounds λ_2.

Here, we give bounds on the eigenvalues derived from simple properties of graphs and manifolds. We will build up to a proof of Cheeger's Inequality, a bound on λ_2 that was first proven on manifolds, but has recently seen widespread use in graph theory.

Theorem 4.4.1. *Let G be a simple connected graph.*

1. *$\lambda_n \leq n$ with equality if and only if the complement $\overline{G}$ is disconnected.*
2. *$\sum_{i=1}^{n} \lambda_i = \sum_{v \in V} d_v = 2|E$*
3. *$\lambda_2 \leq \frac{n}{n-1} \min_{v \in V} d_v$ and $\lambda_n \geq \frac{n}{n-1} \max_{v \in V} d_v$*
4. *$\lambda_n \leq \max_{i \in V}(d_i + m(i))$ where $m(i)$ is the average of the degrees of vertices adjacent to vertex i.*

Proof.

1. From Lemma 4.3.1, the eigenvalues of G are $0, \lambda_2, \ldots, \lambda_n$, those of $\overline{G}$ are $0, n - \lambda_n, \ldots, n - \lambda_n$. The eigenvalues of $\overline{G}$ are nonnegative, so $\lambda_i \leq n$. As shown above, 0 has multiplicity greater than 1 in $\overline{G}$ if and only if $\overline{G}$ is disconnected, so n is an eigenvalue of G if and only if $\overline{G}$ is disconnected.
2. The sum of the eigenvalues of an operator equals its trace, and the trace of $\mathbf{L} = D - A$ is the same as the trace of D, which is the sum of the degree of each vertex: $\sum_{v \in V} d_v$.
3. This result is due to Fielder [32]. For a proof, see Appendix A.2.3.
4. This result is due to Merris [64], building off a result from Anderson and Morley [3]. For a proof, see Appendix A.2.3.

□

Another way of seeing the connection between the Laplacian spectrum and graph connectivity is to observe how they behave as one changes the graph. In particular, if one adds an edge to the graph, the eigenvalues only increase.

Theorem 4.4.2 (Edges Increase Eigenvalues)**.** *Let G be a non-complete graph and (i, j) an edge not in E. Denote by G' the graph G with edge (i, j) added. Then the eigenvalues of G' interlace those of G:*

$$0 = \lambda_1(G) = \lambda_1(G') \leq \lambda_2(G) \leq \lambda_2(G') \leq \lambda_3(G) \leq \cdots \leq \lambda_n(G) \leq \lambda_n(G')$$

The proof of this theorem is included in Appendix A.2.4.[10] It is closely related to Cauchy's Interlace Theorem and Weyl's Theorem, two corollaries of the Courant-Fischer Theorem. It also gives us another way of seeing that the complete graph has the largest eigenvalues.

These types of interlacing results are an active area of research. The theorem above covers the case of edge addition, but for results on vertex addition, edge subdivision, and vertex contraction, we encourage the reader to look at [69].

For manifolds, bounds on the eigenvalues of Δ are often more challenging to prove than their graph counterparts. A well-known result of Lichnerowicz and Obata bounds λ_2 in terms of the Ricci curvature. We will not give a proof, but state it here for readers more familiar with Riemannian geometry.

Theorem 4.4.3 (Lichnerowicz-Obata)**.** *Suppose $\mathcal{M}$ is a compact n-dimensional Riemannian manifold with Ricci curvature satisfying the positive lower bound $Ric(\mathcal{M}) \geq (n-1)K$. Then*

$$\lambda_2(\mathcal{M}) \geq nK$$

with equality if and only if $\mathcal{M}$ is isometric to the sphere $S^n(1)$.

Without the curvature condition of Lichnerowicz-Obata, it is possible for the second eigenvalue of a closed manifolds to be arbitrarily small. In the following example, we construct a dumbbell-shaped object with positive size and arbitrarily small λ_2.

[10] The proof involves background (complex analysis) beyond the expected background of the reader. Nevertheless, we encourage adventurous readers to give it a look!

Example: Cheeger's Dumbbell Consider two spheres of volume V connected by a small cylinder of radius ε and length $2L$. Let f be the function that is 1 on the first sphere, -1 on the second sphere, and linearly decreasing on the cylinder. The gradient of f has norm $1/L$ and is 0 otherwise. Note that $\int_{\mathcal{M}} f dV = 0$. The Rayleigh quotient of f is then

$$\int_{\mathcal{M}} |\nabla f|^2 dV = \frac{L^2}{2V} \text{vol}(C)$$

which goes to 0 as $\varepsilon \to 0$. This quantity upper bounds λ_2, so λ_2 may be made arbitrarily small on a manifold of volume at least $2V$.

4.4.3 Bounds and Boundaries

The Laplacian and its eigenvalues are intimately connected to the boundaries of subsets of the graph. To express this connection, we need a few more definitions.

Let G be a graph and $S \subset V$ be a subset of the vertices of G. We say that the size of the boundary of S is the number of edges between vertices in S and those in $G \setminus S$.

Define the *conductance* of a subset $S \subset V$ of vertices to be the size of its boundary ∂S relative to the size of the subset (or the size of its complement, whichever is smaller):

$$h_G(S) = \frac{|\partial S}{\min(|S|, |G \setminus S|)}$$

Define the conductance of a graph, also called the *Cheeger constant* of G, to be the minimum conductance of any subset:

$$h(G) = \min_{S \subset V} h_G(S)$$

Switching to the manifold case, let $\mathcal{M}$ be a closed n-dimensional manifold. The boundary of an n-dimensional submanifold $S \subset \mathcal{M}$ is $(n-1)$-dimensional. For ease of notation, we write vol$(\cdot)$ to denote the volume of an n-dimensional submanifold and area$(\cdot)$ denote the volume of an $(n-1)$-dimensional region.

Consider a smooth $(n-1)$-dimensional submanifold $B \subset \mathcal{M}$ that divides

$\mathcal{M}$ into two disjoint submanifolds S and T. Let

$$h_{\mathcal{M}}(B) = \frac{\text{area}(B)}{\min(\text{vol}(S), \text{vol}(T))} = \min_{S \subset \mathcal{M}: 0 \leq \text{vol}(S)} \frac{\text{area}(\partial S)}{\min(\text{vol}(S), \text{vol}(M \setminus S))}$$

analogous to h_G above. Also let

$$h(\mathcal{M}) = \min_{S \subset \mathcal{M}} h_{\mathcal{M}}(S)$$

where the minimum is taken over submanifolds S of the form above. We call $h(\mathcal{M})$ the *Cheeger isoperimetric constant* or simply the Cheeger constant of $\mathcal{M}$.

Cheeger's Inequality

Cheeger's inequality is a celebrated result that bounds the conductance of a graph or manifold in terms of λ_2. It is named for geometer Jeff Cheeger, who formulated and proved the result for manifolds.

Theorem 4.4.4 (Cheeger's Inequality for Graphs)**.** *For an unweighted d-regular graph,*

$$h(G) \leq \sqrt{2d\lambda_2}$$

Theorem 4.4.5 (Cheeger's Inequality for Manifolds)**.** *For a closed manifold $\mathcal{M}$,*

$$h(\mathcal{M}) \leq \sqrt{2\lambda_2}$$

We prove both theorems in Appendix A.2.5. The remarkable thing about the two proofs is how similar they are – they are essentially identical!

Measuring Boundaries

We now explore how the Laplacian can be used to measure the size of boundaries.

Starting with the graph case, let $\mathbf{1}_S$ be the characteristic function (i.e. indicator) of a subset $S \subset V$:

$$\mathbf{1}_S(v) = \begin{cases} 1 & v \in S \\ 0 & v \notin S \end{cases}$$

Observe that the size of the boundary may be measured by

$$|\partial S| = \sum_{(i,j) \in E} |\mathbf{1}_S(i) - \mathbf{1}_S(j)| \tag{4.8}$$

because this sum simply counts edges between S and $G \setminus S$.

Turning to the manifold case, let $S \subset \mathcal{M}$ be a n-dimensional submanifold and let $\mathbf{1}_S$ be its characteristic function. The analogous statement to 4.8 above would be

$$|\partial S| = \int_{\mathcal{M}} |\nabla \mathbf{1}_S| \, dV \tag{4.9}$$

but the indicator function is not differentiable on $\partial S \subset \mathcal{M}$, so this expression does not make sense!

If it *did* make sense, we see that it would be consistent with the well-known coarea formula. This formula states that for a Lipschitz function u and an L^1 function g,

$$\int_{\mathcal{M}} g(x) |\nabla u(x)| \, dx = \int_{\mathbb{R}} \left(\int_{u^{-1}(t)} g(x) \, dV_{n-1}(x) \right) dt \tag{4.10}$$

Naively substituting $u = \mathbf{1}_S$ and $g = 1$ into this formula gives Equation 4.9. Of course, $\mathbf{1}_S$ is not Lipschitz, so this substitution is not justified.

It turns out that it *is* possible to formally justify Equation 4.9, but doing so requires the machinery of distribution functions. We informally discuss how this is done in the following section on the Laplacian of the indicator.

The Laplacian of the Indicator

The Laplacian of the indicator function, written $\Delta \mathbf{1}_S$, is a generalization of the derivative of the Dirac delta function. Intuitively, $\Delta \mathbf{1}_S$ is infinitely positive on the inside of the boundary of S, infinitely negative on the outside of the boundary of S, and zero on $S \setminus \partial S$. Formally, it is a distribution function, which is to say that it is only defined in the integrand of an integral, where it integrates to a (generalized) Dirac delta function.

For a function $f : \mathcal{M} \to \mathbb{R}$, integrating $\Delta 1_S f(x)$ gives:

$$\begin{aligned}\int_{\mathcal{M}} \Delta 1_S f(x) dV &= \int_{\mathcal{M}} 1_S \Delta f(x) dV \\ &= \int_S \Delta f(x) dV = \int_S -\mathrm{div}\, \nabla f(x) dV \\ &= \int_{\partial S} (-n \cdot \nabla f)(x) dS\end{aligned}$$

where the first inequality follows from the properties of the Laplacian and the second inequality follows from the divergence theorem. This last integral is called the *surface delta function*, as it generalizes the Dirac delta function. For this reason, the Laplacian of the indicator is also sometimes called the *surface delta prime function*.

In practice, the Dirac delta function is often approximated as the limit of smooth bump functions. In the same way, the Laplacian of the indicator is approximated as the limit of the Laplacian of smooth step functions converging to the indicator function on S.

Example: Smooth Approximation of $\mathbf{\Delta 1}$ on S^1 Since the last two sections were relatively abstract, at this point it may be useful to give a concrete example.

Consider the manifold S^1, viewed as the unit interval $[0, 1]$ with periodic boundary conditions and the canonical metric. Suppose we are interested in calculating the size $|\partial D|$ of a segment D whose length is four-fifths of that of the circle. That is, let D be the region $[0.1, 0.9]$, so $S^1 \setminus D = [0, 0.1) \cup (0.9, 1]$. Figure 4.4.1 (top) shows a diagram of our region.

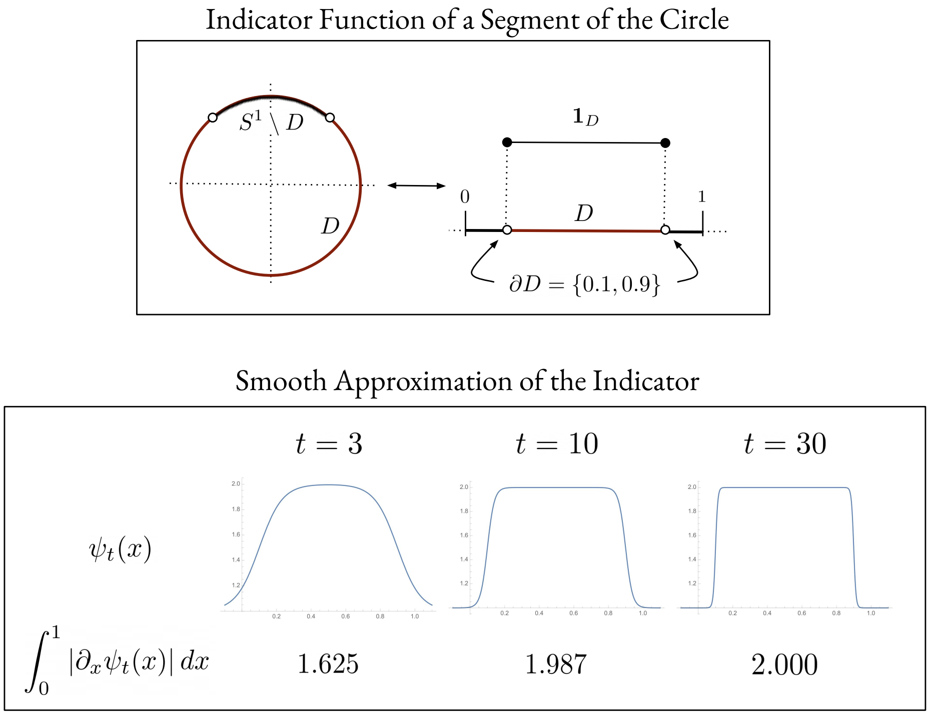
We will create a family of smooth approximations ψ_t, indexed by a parameter t, to the indicator function $\mathbf{1}_D$. We create ψ_t using the sigmoid function

$$\sigma_t(x) = (1 + e^{-x \cdot t})^{-1}$$

which converges to $\mathbf{1}_{x \geq 0}$ as $t \to \infty$. Adding two copies of σ_t and reflecting over the line 0.5 to ensure periodicity, we have

$$\psi_t(x) = \sigma_t(5(1 - x) - 0.5) + \sigma_t(5x - 0.5)$$

Plots of ψ for different values of t are shown in Figure 4.4.1 (bottom). As $t \to \infty$, ψ_i becomes $\mathbf{1}_D$.


Indicator Function of a Segment of the Circle

Smooth Approximation of the Indicator

Figure 4.4.1: Above, an illustration of the indicator function of a segment of a circle. Below, graphs of smooth approximations ψ_t to the indicator for $t = 3, 10, 30$. As t grows large, the integral of $|\partial_x \psi_t(x)|$ approaches $|\partial D| = 2$.

To measure ∂D, we can now compute

$$\int_0^1 |\partial_x \psi_t(x)|\, dx$$

Results of numerical integration using Mathematica for different value of t are displayed in Figure 4.4.1 (bottom). As $t \to \infty$, this quantity approaches 2, which is correct as $|\partial D| = |\{0.1, 0.9\}| = 2$.

4.5 The Heat Kernel

We finish this chapter with a short discussion of the heat equation, the classical motivation for the study of the Laplacian. The heat kernel is the key tool of our main proof in Theorem 5.4.1.

We begin with the manifold variant of the heat equation and then discuss the graph variant.

4.5.1 Manifolds

Let $\mathcal{M}$ be a closed manifold with measure μ. Define the *heat operator* $L : C^2(\mathcal{M}) \times C^1((0,\infty))$ by

$$L = \Delta + \partial_t$$

Let $F(x,t)$ and $f(x)$ be functions on $\mathcal{M} \times (0,\infty)$ and $\mathcal{M}$, respectively. The *heat equation* is the partial differential equation

$$\begin{aligned} Lu(x,t) &= F(x,t) \\ u(x,0) &= f(x) \end{aligned}$$

If $F(x,t) = 0$, we have the *homogenous heat equation*

$$\begin{aligned} Lu(x,t) &= 0 \\ u(x,0) &= f(x) \end{aligned}$$

Theorem 4.5.1. *A solution to the homogeneous heat equation is unique.*

See Appendix A.2.6 for the proof.

A *fundamental solution* to the heat equation is a function $p : \mathcal{M} \times \mathcal{M} \times (0,\infty) \to \mathbb{R}$ that is C^2 on $\mathcal{M} \times \mathcal{M}$ and C^1 on $(0,\infty)$ such that

$$L_y p = 0, \qquad \lim_{t \to 0} p(\cdot, y, t) = \delta_y$$

where δ_y is the Dirac delta function. Fundamental solutions may be shown to be unique and symmetric in x and y.

For $t > 0$, define the *heat propagator* operator $e^{-t\Delta} : L^2(\mathcal{M}) \to L^2(\mathcal{M})$ as

$$e^{-t\Delta} f(x) = \int_{\mathcal{M}} p(x,y,t) f(y) \, d\mu(x)$$

The heat propagator may be thought of as the solution to the heat equation with initial condition $f(x)$. The following theorems state some of its properties; in essence, $e^{-t\Delta}$ behaves as if it were simply an exponentiated function.

Theorem 4.5.2. *The heat propagator satisfies:*

1. $e^{-t\Delta} \circ e^{-s\Delta} = e^{-(s+t)\Delta}$

2. $\left(e^{-\Delta}\right)^t = e^{-t\Delta}$

3. $e^{-t\Delta}$ *is a positive, self-adjoint operator.*

4. $e^{-t\Delta}$ *is compact.*

Theorem 4.5.3. *As* $t \to 0$*,* $e^{-t\Delta} \mapsto Id_{L^2}$*, the identity operator in* $L_2(\mathcal{M})$*.*

Theorem 4.5.4. *As* $t \to \infty$*,* $e^{-t\Delta}$ *converges uniformly in* L^2 *to a constant function (a harmonic function if* $\mathcal{M}$ *is not closed).*

The next theorem reveals the fundamental connection between the heat equation and the Laplacian spectrum.

Theorem 4.5.5 (Sturm-Liouville decomposition)**.** *Denote the eigenvalues and eigenfunctions of the Laplacian* Δ *by* $\lambda_1 \leq \lambda_2 \leq \cdots$ *and* $\phi_1, \phi_2, \ldots,$ *respectively. Then*

$$p(x, y, t) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(x) \phi_i(y)$$

See Appendix A.2.6 for the proof.

4.5.2 Graphs

Having developed our heat operator toolkit on manifolds, we now look at the heat kernel on graphs. In what follows, for ease of notation, we work with the normalized Laplacian $\mathcal{L} = D^{-1/2}\mathbf{L}D^{-1/2}$ rather than $\mathbf{L}$.

For a graph G, we define the heat kernel H_t to match the form $e^{-t\Delta}$:

$$H_t = e^{-t\mathcal{L}}$$

analogously to the Sturm-Liouville decomposition, it may also be written as a sum of outer products,

$$H_t = \phi e^{-t\Lambda} \phi^T = \sum_{i=1}^{|V|} e^{-t\lambda_i} \phi_i \phi_i^T$$

where λ_i and ϕ_i are the eigenvalues and eigenvectors of the $\mathcal{L}$.

For t near 0, $H_t \approx I - \mathcal{L}t$ by a Taylor series expansion; the heat kernel depends only on the graph's local structure. In the limit $t \to 0$, it converges to the identity function, as in Theorem 4.5.3 on manifolds.

Another way of understanding the heat kernel on graphs is to see it as defining a continuous-time random walk. A standard (discrete-time) random walk on G is defined by the random walk matrix P,

$$P_{ij} = \begin{cases} 1/d_i & (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

The entries P_{ij} of P may be regarded as the probability of moving $i \to j$ at any time step. For a distribution v over vertices at time $t = 0$, the entries of $(P^t v)_i$ may be regarded as the probability of being at vertex i after time t.

By a Taylor expansion, the heat kernel H_t may be written as

$$\begin{aligned} H_t = e^{-t\mathcal{L}} &= e^{-t}\left(I + IP + \frac{IP}{2!} + \cdots\right) \\ &= \sum_{k=0}^{\infty} \frac{t^k e^{-t}}{k!} P^k \end{aligned}$$

In this way, it describes a random walk with $Pois(1)$ distributed waiting times.

5
Manifold Regularization

This chapter brings the ideas introduced previously into a unified framework for semi-supervised learning, called manifold regularization.

Manifold regularization is one of a broad set of manifold learning methods based on the *manifold assumption*, which states that the data $\{x\}$ lie on a low-dimensional manifold $\mathcal{M}$ embedded in a higher-dimensional space X.[1]

For example, we expect the space of natural images to be lower dimensional than the pixel space $X = \mathbb{R}^{H \times W \times 3}$ of all ordered sets of $3 \cdot H \cdot W$ pixels; the pixel space is in some sense almost entirely filled with images that look to us like "noise" (in the visual sense). Note that the natural image manifold is clearly nonlinear, as linear interpolating between natural images does not result in more natural images.[2]

[1] The manifold learning hypothesis underlies most popular dimensionality reduction techniques: PCA, Isomaps, Laplacian Eigenmaps, Diffusion maps, local linear embeddings, local tangent space alignment, and many others.

[2] A significant amount of work has gone into trying to identify the intrinsic dimensionality of the image manifold. For instance, [37] suggest that the representations produced by a popular face recognition model have intrinsic dimension 16.

5.0.1 History and Related Work

Introduced by [11] in 2004, manifold regularization gained attention from machine learning practitioners and theoreticians throughout the mid-late 2000s and early 2010s. It was first grounded in a rigorous theory by [10], who justified the use of the data graph Laplacians by proving that, in the limit of infinite data, they converge to data manifold Laplacians. One of the primary objectives of this chapter is to give a clear exposition of this proof using the tools of heat kernels.

A large body of work has emerged around manifold regularization applications and theory in the last decade. Applications include web image annotation, face recognition, human action recognition, and multitask learning [61]. Theoretical analyses have investigated the extent to which the discrete approximations used in manifold regularization (i.e. operators on graphs) conform with the continuous objects that motivate them (i.e. operators on manifolds).

5.0.2 Organization

The organization of this chapter is as follows. We first motivate manifold regularization using the toy example from Chapter 1. Next, we introduce manifold regularization and prove two representer theorems that characterize the solutions to manifold-regularized learning problems. Third, we give examples of algorithms that lie in this framework (Laplacian RLS, Laplacian SVM). Finally, we prove the convergence of the data graph Laplacian to the data manifold Laplacian in the limit of large data.

5.1 Manifold Regularization

Recall the toy example from Chapter 1 (Figure 1.4.1).

When we only consider the labeled data (2 points), our notion of a natural classification function is a straight line, a smooth function in the extrinsic space ($\mathbb{R}^2$). When we add the unlabeled data, our notion of a natural classification function changes to one that is smooth in the intrinsic space (the data manifold).

Now suppose we have a semi-supervised learning problem with N_L labeled examples and N_U unlabeled examples: $S = \{(x_i, y_i)\}_{i=1}^{N_L} \cup \{x_i\}_{i=N_L}^{N_L+N_U}$, for $x_i \in X$ and $y_i \in Y$. We assume the data

x_i are drawn independently from a probability distribution ρ_X supported on a Riemannian manifold $\mathcal{M}$.

Manifold regularization adds a term to the loss function that penalizes functions which are more complex with respect to the intrinsic geometry of the data manifold $\mathcal{M}$:

$$L(f, x, y) = L_{sup}(y, f(x)) + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \gamma_{\mathcal{I}} R_{\mathcal{I}}(f) \tag{5.1}$$

where $\|f\|_{\mathcal{K}}^2$ is standard (extrinsic) Tikhonov regularization term, $R_{\mathcal{I}}(f)$ is a new *intrinsic* regularization term. This intrinsic term captures the intuition that our functions should be smooth *on the manifold*, not just smooth in the extrinsic space.

The constants $\gamma_{\mathcal{K}}$ and $\gamma_{\mathcal{I}}$ determine the strength of extrinsic and intrinsic regularization, respectively. Note that whereas the extrinsic term is data-independent (i.e. it depends only on f), the intrinsic term depends the data (x) by means of the data manifold.

As seen throughout the last chapter, we can measure the smoothness of a function f on $\mathcal{M}$ by the Dirichlet energy, the integral of the Laplacian quadratic form:

$$R_{\mathcal{I}}(f) = \int_{\mathcal{M}} \|f\|_{\mathcal{I}}^2 \, d\rho_X$$

Our objective is then:

$$L(f, x, y) = L_{sup}(y, f(x)) + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \gamma_{\mathcal{I}} \int_{\mathcal{M}} \|\nabla f(x)\| \, d\rho_X(x) \tag{5.2}$$

Clearly, given only finite data, we cannot compute the intrinsic term exactly. The key idea of manifold regularization is to approximate this term by replacing the manifold with a graph approximation.

Suppose we construct a graph G, called a *data graph*, that approximates the data manifold $\mathcal{M}$. For example, we may take G to be the k-nearest neighbors graph (subsection 5.3.1), where each data point x_i is connected by an edge to its k nearest neighbors.

Substituting the Laplacian $\mathbf{L}$ of G for the Laplacian $\Delta_{\mathcal{M}}$ of $\mathcal{M}$, the intrinsic term becomes computable:

$$R_{\mathcal{I}}(f) = \frac{1}{(N_U + N_L)^2} \mathbf{f}(x)^T \mathbf{L} \mathbf{f}(x) \approx \int_{\mathcal{M}} f(x) \Delta_{\mathcal{M}} f(x) \, d\rho_X(x) \tag{5.3}$$

where $\mathbf{f}(x)$ denotes the vector $(f(x_1), \ldots, f(x_n))$. Alternatively, expressed in summation notation, we have:

$$R_{\mathcal{I}}(f) = \frac{1}{(N_U + N_L)^2} \sum_{i=0}^{N} \sum_{r \in N(x_i)} w_{ir}(f(x_i) - f(x_r))^2 \tag{5.4}$$

where w_{ij} is the weight on edge (i, j) if the data graph is weighted, and $w_{ij} = 1$ if the data graph is unweighted.

Substituting $R_{\mathcal{I}}(f)$ back into Equation 5.1 gives the final loss function

$$\sum_{i=0}^{N_L} L_{sup}(y_i, f(x_i)) + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \frac{1}{(N_U + N_L)^2} \mathbf{f}(x)^T \mathbf{L} \mathbf{f}(x) \tag{5.5}$$

for an arbitrary supervised loss function L_{sup}.

In summary, the manifold regularization framework has three steps:

1. Construct a graph from one's data (subsection 5.3.1)
2. Calculate the Laplacian $\mathbf{L}$ of the data graph: $\mathbf{L} = \mathbf{D} - \mathbf{W}$
3. Optimize the regularized objective function:

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \sum_{i=0}^{N_L} L_{sup}(y_i, f(x_i)) + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \frac{1}{(N_U + N_L)^2} \mathbf{f}(x)^T \mathbf{L} \mathbf{f}(x)$$

5.2 Representer Theorems

Now that we can compute our loss function, we are left with the task of optimizing it. Fortunately, as in the case of Tikhonov regularization, we can characterize the form of the optimal solution f^*.

In this section we state and prove two representer theorems: one for the manifold case of Equation 5.2 and one for the graph case of Equation 5.5. We follow the original proofs given in [11].

The standard Representer Theorem (Theorem 3.3.1) expresses the minimizer of a Tikhonov-regularized loss function in terms of the kernel functions evaluated at the data points x. The following manifold regularized extensions are due to [11].

Theorem 5.2.1 (Manifold Regularization Representer Theorem)**.** *Assuming the intrinsic norm $\|\cdot\|_I$ satisfies a smoothness condition*

(Equation 5.10), the minimizer f^* of Equation 5.2 takes the form:

$$f^*(x) = \sum_{i=1}^{N_L} a_i K(x_i, x) + \int_{\mathcal{M}} a(y) K(x, y)\, d\rho_X(y) \tag{5.6}$$

Theorem 5.2.2 (Graph Regularization Representer Theorem)**.** *The minimizer f^* of Equation 5.5 takes the form:*

$$f^*(x) = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x) \tag{5.7}$$

The remainder of this section is dedicated to proving these theorems, beginning with the manifold case.

Idea: The proof is structured as follows. We use an orthogonality argument to show that we can write f^* as the sum of two quantities. The first, corresponding to the first two terms in Equation 5.2, will be a weighted sum of the kernel function at the data points:

$$\sum_{i=1}^{N_L} a_i K(x_i, x)$$

The second, corresponding to the intrinsic term in Equation 5.2, will take the form of a sum $\sum_i a_i e_i$ over basis vectors e_i, where the a_i depend on a differential operator D. A series of lemmas will show that if D is bounded, this sum lies in the span of the integral operator I_K, and so it may be written in the form:

$$\int_{\mathcal{M}} a(y) K(x, y)\, d\rho_X(y)$$

Finally, we will show that D is bounded to complete the proof.

To begin, let $\mathcal{H}_K$ be a RKHS with kernel K and ρ be a distribution supported on a compact manifold $\mathcal{M} \subset X$. Consider the L^2_ρ inner product

$$\langle f, g \rangle_\rho = \int_X f(x) g(x)\, d\rho(x)$$

and let I_K denote the corresponding integral operator

$$(I_K f)(x) = \langle f, k_x \rangle = \int f(y) K(x, y)\, d\rho(y)$$

As noted in 3.2.2, I_K is a compact self-adjoint operator. Denote its eigenfunctions and eigenvalues by $e_1, e_2, \ldots$ and $\lambda_1, \lambda_2, \ldots$, respectively.

The following properties of I_K will prove helpful shortly.

Lemma 5.2.1. *The functions $\sqrt{\lambda_i} e_i$ form an orthonormal basis for $\mathcal{H}_K$.*

Corollary 3. *Any $g \in \mathcal{H}_K$ may be written as $g = \sum_{i=1}^{\infty} b_i e_i$.*

Lemma 5.2.2. *A function $f = \sum_{i=1}^{\infty} a_i e_i$ lies in the image of I_K if and only if*

$$\sum_{i=1}^{\infty} b_i^2 < \infty \tag{5.8}$$

where $b_i = a_i / \lambda_i$.

Proofs of both lemmas are included in Appendix A.3.1.

Next, consider the closure of the span of the kernels of points $x \in \mathcal{M}$, denoted $\mathcal{S}$:

$$\mathcal{S} = \overline{\text{span}\{k_x : x \in \mathcal{M}\}}$$

Note that S with the induced inner product from $\mathcal{H}_K$ is a Hilbert space. Let $\mathcal{H}_{K_\mathcal{M}}$ and $\mathcal{S}_\mathcal{M}$ denote restrictions to $\mathcal{M}$ of $\mathcal{H}_K$ and $\mathcal{S}$, each of which can be seen as Hilbert spaces (with the induced kernel K).

We need two properties of $\mathcal{S}$ and $\mathcal{S}_\mathcal{M}$.

Lemma 5.2.3. $\mathcal{H}_{K_\mathcal{M}} = \mathcal{S}_\mathcal{M}$

Lemma 5.2.4. *The complement of $\mathcal{S}$ is $\mathcal{S}^\perp = \{f \in \mathcal{H} : f(\mathcal{M}) = 0\}$.*

Proofs are included in Appendix A.3.2.

We now return to our learning problem

$$\arg\min_{f \in \mathcal{H}_K} L_{sup}(y, f(x)) + \gamma_\mathcal{K} \|f\|_\mathcal{K}^2 + \gamma_\mathcal{I} \int_\mathcal{M} \|\nabla f(x)\| \, d\rho_X(x) \tag{5.9}$$

We proceed in three steps: (1) we show a solution f exists, (2) we show $f \in \mathcal{S}$, and (3) we show that f has the desired form.

For ease of notation, let H denote the loss we aim to minimize in 5.9.

$$H(f) = L_{sup}(y, f(x)) + \gamma_\mathcal{K} \|f\|_\mathcal{K}^2 + \gamma_\mathcal{I} \|f\|_\mathcal{I}^2$$

where we write $\|f\|_\mathcal{I}^2$ in place of $\int_\mathcal{M} \|\nabla f(x)\| \, d\rho_X(x)$.

Lemma 5.2.5. *A minimizer f^* of Equation 5.9 exists.*

Proof. Consider a ball $\mathcal{B}_r \subset \mathcal{H}_k$ of radius r: $\mathcal{B}_r = \{f \in \mathcal{S} : \|f\|_K \leq r\}$. Since this ball is compact in L^∞, there must exist a minimizer $f_r^* \in \mathcal{B}_r$ of Equation 5.9 in this ball.

The zero function gives us a lower bound on $H(f_r^*)$:

$$H(f_r^*) \leq H(0) = \frac{1}{N_L} \sum_{i=1}^{N_L} L_{sup}(x_i, y_i, 0)$$

If the zero function is a solution, we are done. Otherwise, we obtain a bound on the $\|\cdot\|_K$ term:

$$\|f\|_{\mathcal{K}}^2 \leq \frac{1}{\gamma_{\mathcal{K}}} \left((L_{sup}(y, f(x)) + \gamma_{\mathcal{I}} \|f\|_{\mathcal{I}}^2 \right)) < \frac{1}{N_L \gamma_{\mathcal{K}}} \sum_{i=1}^{N_L} L_{sup}(x_i, y_i, 0)$$

If we keep increasing the radius r of our ball, $H(f)$ must be lower bounded (because the right hand side is fixed). Specifically, the minimizer cannot be found outside the ball of radius $r = \sqrt{\frac{1}{N_L \gamma_{\mathcal{K}}} \sum_{i=1}^{N_L} L_{sup}(x_i, y_i, 0)}$.

Therefore there exists a solution f^*.

Also, if V is convex then the full objective is convex and the solution is unique. □

Lemma 5.2.6. *If the intrinsic norm $\|\cdot\|_{\mathcal{I}}$ satisfies the following smoothness condition:*

$$f|_{\mathcal{M}} = g|_{\mathcal{M}} \implies \|f\|_I = \|g\|_I \qquad \forall f, g \in \mathcal{H}_K \tag{5.10}$$

Then the solution f^* of Equation 5.9 lies in $\mathcal{S}$.

Proof. Let $f \in \mathcal{H}_K$. Decompose f into the orthogonal projections $f = f_{\mathcal{S}} + f_{\mathcal{S}^\perp}$. By Lemma 5.2.4, $f_{\mathcal{S}^\perp} = 0$ on $\mathcal{M}$. Then $(f - f_{\mathcal{S}}) = 0$ on $\mathcal{M}$, so for the intrinsic norm:

$$\|f\|_I^2 = \|f_{\mathcal{S}}\|_I^2$$

For the extrinsic norm, we have

$$\|f\|_K^2 = \|f_{\mathcal{S}}\|_K^2 + \|f_{\mathcal{S}^\perp}\|_K^2$$

which implies

$$\|f\|_K^2 \geq \|f_{\mathcal{S}}\|_K^2$$

This shows that $f^* \in \mathcal{S}$, because if f^* had any component orthogonal to $\mathcal{S}$, this component would contribute strictly positively to the expression in Equation 5.9. □

From now on, we will assume that $\|\cdot\|_I$ satisfies the smoothness condition (5.10).

We have finally built up to the main result.

Theorem 5.2.3. *The minimizer f^* of*

$$H(f) = \frac{1}{N_L}\sum_{j=1}^{N_L} L_{sup}(y_j, f(x_j)) + \gamma_{\mathcal{K}}\|f\|_{\mathcal{K}}^2 + \gamma_{\mathcal{I}}\|f\|_{\mathcal{I}}^2 \tag{5.11}$$

takes the form:

$$f^*(x) = \sum_{i=1}^{N_L} a_i K(x_i, x) + \int_{\mathcal{M}} a(y)K(x,y)\,d\rho_X(y) \tag{5.12}$$

Proof. By Lemma 5.2.5, a minimizer $f^*(x)$ exists. By Lemma 5.2.6, $f^*(x) \in \mathcal{S}$, the closure of kernel functions centered at points in $\mathcal{M}$. By Lemma 5.2.1, we can write $f^* = \sum_{i=1}^{\infty} a_i e_i$, where $\{e_i\}$ are the basis formed by the eigenvectors of the integral operator I_K, which we defined above as $I_K(f) = \sum_{\mathcal{M}} f(y)K(x,y)\,d\rho_X(y)$.

We will show that f^* decomposes into two terms, the first of which is a finite sum of kernel functions at the data points x_i, and the second of which is lies in the image of I_K and so may be written as $\sum_{\mathcal{M}} a(y)K(x,y)\,d\rho_X(y)$ for some function a.

To begin, we plug $f^* = \sum_{i=1}^{\infty} a_i e_i$ into H:

$$H(f^*) = \frac{1}{N_L}\sum_{j=1}^{N_L} L_{sup}\left((y_j, \sum_{i=1}^{\infty} a_i e_i(x_j)\right)) + \gamma_{\mathcal{K}}\left\|f\sum_{i=1}^{\infty} a_i e_i\right\|_{\mathcal{K}}^2 + \gamma_{\mathcal{I}}\left\|f\sum_{i=1}^{\infty} a_i e_i\right\|_{\mathcal{I}}^2$$

We differentiate with respect to a_k and set the result to 0:

$$0 = \frac{\partial H(f^*)}{\partial a_k} = \frac{1}{N_L}\sum_{j=1}^{N_L} e_k(x_j)\partial_{(2)}L_{sup}(y_j, f^*(x_j)) + 2\gamma_{\mathcal{K}}\frac{a_k}{\lambda_k} + \gamma_{\mathcal{I}}\langle (D+D^*)f, e_k\rangle$$

where D is a differential operator, D^* is its adjoint, and $\partial_{(2)}$ is the partial with respect to the second input of L_{sup}. Note the two terms above corresponding to the norms hold because

$$\frac{\partial H(f^*)}{\partial a_k} \left\| f \sum_{i=1}^{\infty} a_i e_i \right\|_{\mathcal{K}}^2 = \frac{\partial H(f^*)}{\partial a_k} \sum_{i=1}^{\infty} \frac{a_i^2}{\lambda_i} = 2\frac{a_k}{\lambda_k} \qquad \text{and}$$

$$\frac{\partial H(f^*)}{\partial a_k} \left\| f \sum_{i=1}^{\infty} a_i e_i \right\|_{\mathcal{K}}^2 = \langle Df, e_k \rangle + \langle f, De_k \rangle = \langle (D + D^*) f, e_k \rangle$$

Solving the equation above for a_k yields

$$a_k = -\frac{\lambda_k}{2\gamma_{\mathcal{K}}} \left(\gamma_{\mathcal{I}} \frac{1}{N_L} \sum_{j=1}^{N_L} e_k(x_j) \partial_{(2)} L_{sup}(y_j, f^*(x_j)) + \langle (D + D^*) f, e_k \rangle \right)$$

We can plug this expression back into $f^* = \sum_{i=1}^{\infty} a_i e_i$ to give

$$f^*(x) = -\frac{1}{2\gamma_{\mathcal{K}} N_L} \sum_{j=1}^{N_L} \sum_{k=1}^{\infty} \lambda_k e_k(x_j) e_k(x) \partial_{(2)} L_{sup}(y_j, f^*(x_j)) - \frac{\lambda_k}{2\gamma_{\mathcal{K}}} \sum_{k=1}^{\infty} \lambda_k \langle (D + D^*) f, e_k \rangle e_k$$

Using the fact that $K(x, y) = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(y)$, we have:

$$f^*(x) = \underbrace{-\frac{1}{2\gamma_{\mathcal{K}} N_L} \sum_{j=1}^{N_L} K(x, x_j) \partial_{(2)} L_{sup}(y_j, f^*(x_j))}_{\text{sum of kernels at data } x_j} - \underbrace{\frac{\lambda_k}{2\gamma_{\mathcal{K}}} \sum_{k=1}^{\infty} \lambda_k \langle (D + D^*) f, e_k \rangle e_k}_{\text{this is in the image of } I_K}$$

The first term above takes our desired form. By Lemma 5.2.2, the second term above is in the image of I_K if and only if:

$$\sum_{k=1}^{\infty} \frac{(\lambda_k \langle (D + D^*) f, e_k \rangle)^2}{\lambda_k^2} = \sum_{k=1}^{\infty} \langle (D + D^*) f, e_k \rangle^2$$

is bounded. Lemma 5.2.7 below shows that D is bounded, implying that $D + D^*$ is bounded and so the expression above is bounded. Given this result, the second term above is in the image of I_K, and so takes the form $\int_{\mathcal{M}} g(y) K(x, y) \, d\rho_X(y)$.

Therefore

$$f^*(x) = \sum_{i=1}^{N_L} a_i K(x_i, x) + \int_{\mathcal{M}} a(y) K(x, y)\, d\rho_X(y)$$

for some real numbers a_i and some function a. □

To complete the proof, all that remains is to show that D is bounded. To do so, we have to be a bit more specific about the geometry of our manifold. Let $\mathcal{M}$ be a boundaryless manifold with measure ρ, $D \in C^\infty$ a differential operator, and $K(x, y)$ a kernel with at least $2k$ derivatives.

Lemma 5.2.7. *$D : \mathcal{S} \to L^2_\rho$ is a bounded operator.*

Proof. We show D is bounded on $\mathcal{H}_K$. Note that the integral operator I_K defined above is compact (and so bounded). As a result, $I_K D$ is bounded, and (by taking the adjoint and composing with D^*) we have that $DI_KD^* : L^2_\rho \to L^2_\rho$ is bounded.

Consider the square root $I_K^{1/2}$ of I_K. As seen by the eigenvalues of $I_K^{1/2}$ or the relation $I_K^{1/2} \circ I_K^{1/2} = I_K$, this operator is positive and adjoint. As seen above, $I_K^{1/2} : \mathcal{H}_K \to L^2_\rho$ is an isometry, so any $g \in \mathcal{H}_K$ may be written as $I_K^{1/2} f$ for some $f \in L^2_\rho$. Then $\|f\|_{L^2_\rho} = \|g\|_K$. We now have

$$\|Dg\|_{L^2_\rho} = \left\|DI_K^{1/2} f\right\|_{L^2_\rho} \le \left\|DI_K^{1/2} f\right\|_{L^2_\rho} \|f\|_{L^2_\rho} = \left\|DI_K^{1/2} f\right\|_{L^2_\rho} \|g\|_K \tag{5.13}$$

Finally, we bound $DI_K^{1/2}$. Let $\varepsilon > 0$ be arbitrary. There exists $f \in L^2_\rho$ such that $\|f\|_{L^2_\rho}$ and

$$\left\|DI_K^{1/2}\right\|_{L^2_\rho}^2 = \left\|I_K^{1/2} D^*\right\|_{L^2_\rho}^2 \le \langle I_K^{1/2} D^* f, I_K^{1/2} D^* f\rangle_{L^2_\rho} = \langle DI_K D^*, f\rangle_{L^2_\rho} \le \|DI_K D^*\| \, \|f\|^2$$

Now $\|f\|^2 \le (1+\varepsilon)^2$ and $\|DI_KD^*\|$ is bounded, so $\left\|DI_K^{1/2}\right\|_{L^2_\rho}^2$ is bounded.

Returning to Equation 5.13, we see:

$$\|Dg\|_{L^2_\rho} \le \left\|DI_K^{1/2} f\right\|_{L^2_\rho} \|g\|_K \le C \cdot \|g\|_K$$

for some constant C. Therefore D is a bounded operator $\mathcal{S} \to L^2_\rho$. □

With this result, our proof of Theorem 5.2.2 is complete.

Fortunately, the proof of the discrete manifold regularization theorem is significantly simpler. It parallels the orthogonality argument from the original representer theorem.

Theorem 5.2.4 (Theorem 5.2.2)**.** *The minimizer f^* of*

$$H(f) = \frac{1}{N_L}\sum_{j=1}^{N_L} L_{sup}(y_j, f(x_j)) + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \frac{\gamma_{\mathcal{I}}}{(N_L + N_U)^2}\boldsymbol{f}^T \boldsymbol{L}\boldsymbol{f} \tag{5.14}$$

takes the form:

$$f^*(x) = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x) \tag{5.15}$$

Proof. Suppose f is a minimizer of Equation 5.14. Let S be the subspace spanned by the kernel functions K_{x_i} on the data $\{x_i\}_{i=1}^{N_L+N_U}$, in other words the functions that may be written in the form $\sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$ for some coefficients a_i.

Write $f = f_S + f_{S^\perp}$, where f_S and $f_{S^\perp}$ are orthogonal projections onto S and $S^\perp$. Our goal is to show that $f_{S^\perp} = 0$, as then f takes the form $f(x) = f_S(x) = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$.

By the reproducing property, we see that the value of f on a data point x_i does not depend on $f_{S^\perp}$:

$$f(x_i) = \langle f, K_{x_i}\rangle = \langle f_S, K_{x_i}\rangle + \langle f_{S^\perp}, K_{x_i}\rangle = \langle f_S, K_{x_i}\rangle$$

Examining Equation 5.15, the first and third components of $H(f)$ only depend on f evaluated at the data points. Therefore $H(f)$ and $H(f_S)$ differ only on the second component:

$$H(f) - H(f_S) = \|f\|_K^2 - \|f_S\|_K^2 = \|f_{S^\perp}\|_K^2$$

If f is a minimizer of H, this difference cannot be positive, so:

$$\|f_{S^\perp}\|_K^2 \leq 0 \implies \|f_{S^\perp}\|_K^2 = 0 \implies f_{S^\perp} = 0$$

Therefore $f = f_S \in S$ and f takes the form

$$f(x) = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$$

□

Whereas the manifold-based representer theorem is exclusively of theoretical interest, this graph-based version enables us to compute solutions to manifold regularized learning problems. We give two examples of such algorithms below.

5.3 Algorithms

In general, to solve a manifold regularized learning problem, we solve for a function in the form given by the representer theorem

$$f(x) = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$$

by optimizing the parameters a_i, usually using gradient-based optimization methods.

Laplacian Regularized Least Squares (Lap-RLS) Lap-RLS corresponds to a least squares loss function on the supervised data, $L_{sup}(f(x), y) = (f(x) - y)^2$. Our objective is then

$$f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{N_L} \sum_{j=1}^{N_L} (f(x_i) - y_i)^2 + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \frac{\gamma_{\mathcal{I}}}{(N_L + N_U)^2} \mathbf{f}^T \mathbf{L} \mathbf{f}$$

By the representer theorem, our minimizer takes the form $f^* = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$. At this point, we would usually use gradient descent on the a_i, but in this case we are able to give a closed form.

To simplify notation, define:

- $\mathbf{a} = (a_1, \ldots, a_{N_L+N_U}) \in \mathbb{R}^{(N_L+N_U)}$ to be the vector of coefficients a_i
- $K = (K(x_i, x_j))_{i,j=1}^{N_L+N_U} \in \mathbb{R}^{(N_L+N_U)\times(N_L+N_U)}$ to be the kernel matrix (or Gram matrix) on the labeled and unlabeled data
- $Y = (y_1, \ldots, y_{N_L}, 0, \ldots, 0) \in \mathbb{R}^{N_L+N_U}$ to be the label vector on the labeled data and 0 on the unlabeled data
- $J = \text{diag}(1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{(N_L+N_U)\times(N_L+N_U)}$ to be the matrix with 1s on the diagonal entries corresponding to the labeled data and

0 elsewhere.

Plugging in $f^* = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$, our objective is:

$$\mathbf{a}^* = \arg \min_{\mathbf{a}\in\mathbb{R}^{N_L+N_U}} \frac{1}{N_L}(Y - JK\mathbf{a})^T(Y - JK\mathbf{a}) + \gamma_{\mathcal{K}}\mathbf{a}^T K\mathbf{a} + \frac{\gamma_{\mathcal{I}}}{(u+l)^2}\mathbf{f}^T\mathbf{L}\mathbf{f}$$

Taking a derivative and solving for $\mathbf{a}^*$ gives:

$$\mathbf{a}^* = \left(JK + \gamma_{\mathcal{K}} I + \tfrac{\gamma_{\mathcal{I}}}{(N_L+N_U)^2} LK\right)^{-1} Y$$

This is the same as the well-known solution $w^* = (K + \gamma_{\mathcal{K}} I)^{-1} Y$ of the standard RLS problem, with an added term corresponding to the intrinsic norm.

Laplacian Support Vector Machines (Lap-SVM) Lap-SVM corresponds to a hinge loss on the supervised data, $L_{sup}(f(x), y) = \max(0, 1 - yf(x)) = (1 - yf(x))_+$ where $y \in \{-1, 1\}$. Our objective is then

$$f^* = \arg \min_{f\in\mathcal{H}} \frac{1}{N_L} \sum_{j=1}^{N_L} \max(0, 1 - y_i f(x_i)) + \gamma_{\mathcal{K}} \|f\|_{\mathcal{K}}^2 + \frac{\gamma_{\mathcal{I}}}{(N_L+N_U)^2}\mathbf{f}^T\mathbf{L}\mathbf{f}$$

Again by the representer theorem, $f^* = \sum_{i=1}^{N_L+N_U} a_i K(x_i, x)$ and we are looking for:

$$\mathbf{a}^* = \arg \min_{\mathbf{a}\in\mathbb{R}^{N_L+N_U}} \left(\frac{1}{N_L} \sum_{j=1}^{N_L} \max\left(0, 1 - y_i \left(\sum_{i=1}^{N_L+N_U} a_i K(x_i, x)\right)\right) \right.$$
$$\left. + \gamma_{\mathcal{K}}\mathbf{a}^T K\mathbf{a} + \frac{\gamma_{\mathcal{I}}}{(N_L+N_U)^2}\mathbf{a}^T KLK\mathbf{a} \right)$$

A Note on Complexity The primary difficulty with using Lap-RLS, Lap-SVM and similar algorithms in practice is the computational complexity of working with the kernel matrix K, a dense $(N_L + N_U) \times (N_L + N_U)$ matrix. In Lap-RLS, for example, the matrix inversion takes $O((N_L + N_U)^3)$ time, which is infeasible for datasets containing millions of unlabeled examples.

Developing sparse and computationally tractable approximations for the types of objective functions seen above is an active area of research. In fact, it is most active in the Gaussian processes research community, which faces the same challenge of inverting large kernel matrices in Gaussian process regression.

An Note on the Hessian Thus far, almost all our work has been based on the Laplacian operator. A somewhat less popular but still notable theory has arisen in parallel that substitutes the Hessian for the Laplacian. Changing from a Laplacian-regularized loss function to a Hessian-regularized is as simple as changing the quadratic form $f^T \mathcal{L} f$ to $f^T \mathcal{H} f$.

Theoretically, whereas the Laplacian corresponds to the Dirichlet Energy, the Hessian corresponds to the Eells Energy:

$$E_{Eells}(f) = \int_{\mathcal{M}} \|\nabla_a \nabla_b f\|^2_{T^*_x \mathcal{M} \otimes T^*_x \mathcal{M}} \, dV(x)$$

Manipulating this expression into normal coordinates yields the Frobenius norm of the Hessian of f:

$$\mathbb{R}(f) = \sum_{i=1}^{N} \sum_{r,s=1}^{m} \left(\frac{\partial^2 f}{\partial x_r \partial x_s}(x_i) \right)^2$$

However, the second-order nature of the Hessian is a double-edged sword. While it gives the operator the desirable properties mentioned above, it makes the Hessian difficult to compute. To get around this, [53] introduce a sparse matrix approximation $\mathbf{B}$ by fitting a quadratic function to the data points. This approximation yields an objective function almost identical to that of Laplacian-based manifold regularization:

$$L(f, x, y) = L_{sup}(y, f(x)) + \gamma_{\mathcal{K}} \|f\|^2_{\mathcal{K}} + \gamma_{\mathcal{I}} \mathbf{f}^T \mathbf{B} \mathbf{f}$$

where $\mathbf{B}$ is analogous to $\mathbf{F}$ in Equation 5.11.

5.3.1 Data Graphs

Thus far, we have glossed over the first step of manifold learning algorithms: constructing a graph from the data. Here, we briefly give a summary of different types of data graphs. In all cases, the data graph $G = (V, E)$ is undirected and its vertices V correspond to the observed

Type	*Sparse*	*Connected*	*Construction Time*
k-Nearest Neighbors	✓	✓	*Varies*
ε-Neighbors	✓	✗	*Varies*
Gaussian	✗	✓	$O(n^2)$
b-Matching	✓	✓	$O(dn^3)$

Table 5.3.1: A comparison of different graph construction methods. Note that the running time for k-nearest neighbors and ε-neighbors methods depends on the neighbor-finding algorithm chosen. Usually, a fast, approximate algorithm is chosen rather than an exact algorithm. It is also possible to improve the speed of b-matching graph construction with loopy belief propagation.

data $\{x_i\}_{i=1}^N$.

Common data graphs include:

- *k-Nearest-Neighbors Graph*: An edge is created between each data point x and the k other points closest to x (nearest neighbors) according to some distance function d. This graph is sparse and connected.

- *ε-Neighbors Graph*: An edge is created between all pairs (x, x') of data points with distance less than ε according to a distance function d. Each edge has weight 1. This graph is sparse, but it may be disconnected.

- *Gaussian-Weighted Graph*: A fully-connected weighted graph is constructed using Gaussian edge weights: $w_{ij} = e^{-\frac{(x_i-x_j)^2}{\sigma^2}}$ for some $\sigma^2 > 0$. This graph turns out to have attractive theoretical properties, but unlike the other graphs here it is dense, so it is computationally difficult to work with.

- *b-Matching Graph*: A graph is obtained by solving a maximum weight matching problem: $\min_w \sum_{w_{ij}d(x_i,x_j)}$ subject to the constraints that $w_{i,j}$ is binary, symmetric, and b-regular (i.e. every node has exactly b edges). The solution is sparse, connected, and b-regular by construction. It has been found to perform well on small to medium-sized datasets, but solving the matching problem can take $O(dn^3)$ time.

Recent research on graph construction includes methods based on random walks [70], adaptive coding [88], signal representation [29], and ensembles of different types of graphs [4].

5.4 Convergence of the Graph Laplacian

This section is dedicated to proving that the Laplacian matrix $\mathbf{L}$ on an (exponentially-weighted) data graph converges to the Laplacian $\Delta_{\mathcal{M}}$ on the data manifold. This result is fundamental to all Laplacian-based manifold learning techniques, as it justifies the replacement of $\Delta_{\mathcal{M}}$ with $\mathbf{L}$ in Equation 5.3, which makes optimization tractable.

Let $\mathcal{M}$ be a compact k-dimensional manifold embedded in $\mathbb{R}^N$. Let $S = \{x_i\}_{i=1}^n$ for x_i sampled i.i.d. from the uniform distribution on $\mathcal{M}$ (that is, the distribution $\rho(x) = 1/\text{vol}(\mathcal{M})$ for $x \in \mathcal{M}$).

Notation: In this chapter, we will use n to denote the number of data points and N to denote the dimension of the ambient space. n here corresponds to $N_L + N_U$ above.

The following result is due to [10].

Theorem 5.4.1 (Convergence of the Graph Laplacian)**.** *Fix a function $f \in C^\infty(\mathcal{M})$, a point $z \in \mathcal{M}$, and a constant $a > 0$. Set $t_n = n^{1/(k+2+a)}$. Then*

$$\lim_{n\to\infty} \frac{1}{t_n(4\pi t_n)^{\frac{k}{2}}} \boldsymbol{L}_n^{t_n} f(z) = \frac{1}{vol(\mathcal{M})} \Delta_{\mathcal{M}} f(z)$$

where the limit holds in probability.

The proof has three steps. The first two steps show that $\mathbf{L}^t$ converges to $\Delta_{\mathcal{M}}$ as $t \to 0$ using the heat operator. The final step shows that $\frac{1}{n}\mathbf{L}_n^t$ converges to $\mathbf{L}^t$ as $n \to \infty$ using Hoeffding's inequality.

The key idea of the proof is that if we construct a weighted graph from the data points $\{x_i\}_{i=1}^n$ with Gaussian edge weights, we can associate its (discrete) Laplacian with the (continuous) heat kernel on $\mathcal{M}$.

To begin, let $G = (V, E)$ be a fully-connected weighted graph on $|V| = n$ vertices, with each vertex corresponding to a data point $x_i \in S$. Assign to each edge $(i, j) \in G$ the weight

$$w_{ij} = e^{\frac{\left\|x_i - x_j\right\|^2}{4t}}$$

where $t > 0$. Note that G varies with the number of nodes n and the parameter t.

Consider the Laplacian matrix of G, which we write as $\mathbf{L}_n^t$:

$$\begin{aligned}\mathbf{L}_n^t f(x_i) &= f(x_i)\sum_{j=1}^{n} w_{ij} - \sum_{j=1}^{n} f(x_j)w_{ij} \\ &= f(x_i)\sum_{j=1}^{n} e^{\frac{\left\|x_i-x_j\right\|^2}{4t}} - \sum_{j=1}^{n} f(x_j)e^{\frac{\left\|x_i-x_j\right\|^2}{4t}}\end{aligned}$$

We may extend $\mathbf{L}_n^t$ to a linear operator on functions defined on the ambient space of points $x \in \mathbb{R}^N$:

$$\mathbf{L}_n^t f(x) = f(x)\sum_{j=1}^{n} e^{-\frac{\left\|x-x_j\right\|^2}{4t}} - \sum_{j=1}^{n} f(x_j)e^{-\frac{\left\|x-x_j\right\|^2}{4t}}$$

The continuous analogue of this operator, which we denote $\mathbf{L}^t$, generalizes the expression from a discrete set of points x_j to a measure ρ:

$$\begin{aligned}\mathbf{L}^t f(x) &= f(x)\int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4t}}\, d\rho(y) - \int_{\mathcal{M}} f(y)e^{-\frac{\|x-y\|^2}{4t}}\, d\rho(y) \\ &= \int_{\mathcal{M}} (f(x) - f(y))e^{-\frac{\|x-y\|^2}{4t}}\, d\rho(y)\end{aligned}$$

We would like to show that as $t \to 0$, after appropriate scaling, $\mathbf{L}^t$ converges to $\Delta_{\mathcal{M}}$. More precisely:

Lemma 5.4.1. *Fix $z \in \mathcal{M}$. Then:*

$$\lim_{t\to 0} \frac{1}{t(4\pi t)^{k/2}} \boldsymbol{L}^t f(z) = \frac{1}{vol(\mathcal{M})}\Delta_{\mathcal{M}} f(z)$$

The proof of this lemma has two sub-parts. First, we restrict our attention to an open ball B around $z \in \mathcal{M}$ and perform an exponential change of coordinates. This coordinate transformation reduces our computations to computations in $\mathbb{R}^k$. Second, we show that our (transformed) integral involving $\mathbf{L}^t$ converges to the Laplacian in $\mathbb{R}^N$.

Essentially, the idea is that since the manifold is locally Euclidean, we can restrict our attention to a local space and then prove our result using properties of Gaussians integrals in $\mathbb{R}^N$.

Step 1: Exponential Change of Coordinates

First, we restrict to a local ball $B \subset \mathbb{R}^N$ around z.

Lemma 5.4.2. *For any $f \in L^\infty(\mathcal{M})$ and any $a > 0$,*

$$\lim_{t \to 0} \left| \int_{\mathcal{M}} f(y) e^{-\frac{\|z-y\|^2}{4t}} \, d\rho(y) - \int_B f(y) e^{-\frac{\|z-y\|^2}{4t}} \, d\rho(y) \right| = o(t^a) \text{ as } t \to 0 \tag{5.16}$$

Proof. Denote by B^c the complement of B in $\mathcal{M}$. We write $\rho(B^c)$ for the measure of B^c under ρ.

Let $d = \inf x \in B^c \|p - x\|^2$. We see that on B^c, the integrand $f(y)e^{-\frac{\|z-y\|^2}{4t}}$ is bounded by $\sup_{x \in B^c} (|f(x)|) \, e^{-\frac{d^2}{4t}}$.

The the difference between the integrals over $\mathcal{M}$ and B is bounded by:

$$\left| \int_{\mathcal{M}} f(y) e^{-\frac{\|z-y\|^2}{4t}} \, d\rho(y) - \int_B f(y) e^{-\frac{\|z-y\|^2}{4t}} \, d\rho(y) \right| \leq \rho(B^c) \cdot \sup_{x \in B^c} (|f(x)|) \, e^{-\frac{d^2}{4t}}$$

Since $\rho(B^c)$ and $\sup_{x \in B^c} (|f(x)|)$ are constant, this decreases exponentially with t, so it is $o(t^a)$ for any $a > 0$. □

It follows from this lemma that

$$\mathbf{L}^t f(p) - \int_B (f(p) - f(y)) e^{-\frac{\|p-y\|^2}{4t}} \, d\rho(y) = o(t^a) \tag{5.17}$$

We now transform to the canonical coordinates of $\mathcal{M}$, as described in Section 4.1.1.

For $x \in B$, the ball around z, consider a transformation $y = \exp_z(x)$ under the canonical map $\exp_z : T_z\mathcal{M} \to \mathcal{M}$. As discussed earlier, $\exp_z$ is a diffeomorphism from a local neighborhood B of 0 in the tangent space $T_z\mathcal{M}$ to a local neighborhood $\tilde{B}$ of z in the manifold $\mathcal{M}$, with the property that $\exp_z(0) = z$. Since the tangent space $T_z\mathcal{M}$ can be naturally identified with $\mathbb{R}^k$, we consider $\exp_z$ to be a diffeomorphism from a neighborgood $B \subset \mathcal{M}$ around z to a neighborhood $\tilde{B} \subset \mathbb{R}^k$ around 0.

Given a function $f : \mathcal{M} \to \mathbb{R}$, let $\tilde{f} : \mathbb{R}^k \to \mathbb{R}$ denote its local coordinate transformation about z. That is, $\tilde{f} = f \circ \exp_z$, or

$$\tilde{f}(x) = f(\exp_z(x))$$

which makes sense because for $y \in B \subset \mathcal{M}$, we can write $y = exp_z(x)$ for $x \in \mathbb{R}^N$. A change of coordinates shows

$$\Delta_{\mathcal{M}} f(p) = \Delta_{\mathbb{R}^k} \tilde{f}(0)$$

Also note that for any $z \in B \subset \mathcal{M}$ and $x \in \mathbb{R}^k$,

$$0 \leq \|x\|^2_{\mathbb{R}^K} - \|z - \exp_z(x)\|^2_{\mathbb{R}^N} \leq C \cdot \|x\|^4_{\mathbb{R}^K}$$

We can now write the integral in Equation 5.17, which approximates $\mathbf{L}^t$, as

$$\begin{aligned}\int_B (f(p) - f(y)) e^{-\frac{\|z-y\|^2}{4t}}\, d\rho(y) &= \frac{1}{\mathrm{vol}(\mathcal{M})} \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x)) \sqrt{\det(g_{ij})}\, e^{-\frac{\|\exp_p(x) - y\|^2}{4t}}\, dx \\ &= \frac{1}{\mathrm{vol}(\mathcal{M})} \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x))(1 + O(\|x\|^2)) e^{-\frac{\|x\|^2 - g(x)}{4t}}\, dx\end{aligned}$$

which is a standard integral over $\mathbb{R}^n$. We denote a scaled version of this integral by L:

$$L = \frac{1}{t(4\pi t)^{k/2}} \frac{1}{\mathrm{vol}(\mathcal{M})} \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x))(1 + O(\|x\|^2)) e^{-\frac{\|x\|^2 - g(x)}{4t}}\, dx$$

For ease of future notation, let $\gamma_t = \frac{1}{t(4\pi t)^{k/2}} \frac{1}{\mathrm{vol}(\mathcal{M})}$, so

$$L = \gamma_t \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x))(1 + O(\|x\|^2)) e^{-\frac{\|x\|^2 - g(x)}{4t}}\, dx \tag{5.18}$$

Step 2: Convergence in $\mathbb{R}^k$

Consider $e^{-\frac{\|x\|^2 - g(x)}{4t}}$ in $\mathbb{R}^k$. By a Taylor Series, we may approximate it as

$$e^{-\frac{\|x\|^2 - g(x)}{4t}} = e^{-\frac{\|x\|^2}{4t}} \cdot e^{\frac{g(x)}{4t}} = e^{-\frac{\|x\|^2}{4t}} \left(1 + O\left(\frac{g(x)}{4t} e^{\frac{g(x)}{4t}}\right)\right)$$

We can then rewrite L from 5.18 as:

$$\begin{aligned}L &= \gamma_t \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x))(1 + O(\|x\|^2)) e^{-\frac{\|x\|^2}{4t}} \left(1 + O\left(\frac{g(x)}{4t} e^{\frac{g(x)}{4t}}\right)\right) dx \\ &= A_t + B_t + C_t\end{aligned}$$

where

$$
\begin{aligned}
A_t &= \gamma_t \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x)) e^{-\frac{\|x\|^2}{4t}}\, dx \\
B_t &= \gamma_t \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x)) e^{-\frac{\|x\|^2}{4t}} O(\|x\|^2)\, dx \\
C_t &= \gamma_t \int_{\tilde{B}} (\tilde{f}(0) - \tilde{f}(x)) e^{-\frac{\|x\|^2}{4t}} (1 + O(\|x\|^2))\, O\left(\tfrac{g(x)}{4t} e^{\frac{g(x)}{4t}} \right) dx
\end{aligned}
$$

We now show that in the limit $t \to 0$,

$$A_t = \tfrac{\Delta_{\mathcal{M}} f(p)}{\operatorname{vol}(\mathcal{M})}, \quad B_t = 0, \quad C_t = 0,$$

For A_t, by the Taylor expansion of $\tilde{f}(x) - \tilde{f}(0)$,

$$
\begin{aligned}
A_t &= -\gamma \int_{\tilde{B}} \left(x \nabla f(0) + \frac{x^T H x}{2} + O(\|x\|^3) \right) e^{-\frac{\|x\|^2}{4t}}\, dx \\
&= -\gamma \int_{\tilde{B}} x e^{-\frac{\|x\|^2}{4t}} \nabla f(0)\, dx - \operatorname{vol}(\mathcal{M}) \sum_{i,j=1}^{k} \tfrac{1}{t(4\pi t)^{k/2}} \int_{\tilde{B}} \tfrac{1}{2} x_i x_j H_{ij} e^{-\frac{\|x\|^2}{4t}}\, dx \\
&\quad - \gamma \int_{\tilde{B}} O(\|x\|^3\, e^{-\frac{\|x\|^2}{4t}}\, dx
\end{aligned}
$$

where H is the Hessian of $\mathbf{f}$ at 0. By the properties of Gaussian integrals in $\mathbb{R}^k$:

$$(1) \qquad \int_{\tilde{B}} x e^{-\frac{\|x\|^2}{4t}}\, dx = 0$$

$$(2) \qquad \int_{\tilde{B}} x_i x_j e^{-\frac{\|x\|^2}{4t}}\, dx = 0 \qquad \text{if } i \neq j \text{ and} \qquad \lim_{t \to 0} \tfrac{1}{t(4\pi t)^{k/2}} \int_{\tilde{B}} x_i^2 e^{-\frac{\|x\|^2}{4t}}\, dx = 2$$

$$(3) \qquad \tfrac{1}{t(4\pi t)^{k/2}} \int_{\tilde{B}} \|x\|^3\, e^{-\frac{\|x\|^2}{4t}}\, dx = O(t^{1/2})$$

Applying these properties to the terms of A_t, we have

$$A_t = 0 + \frac{1}{\operatorname{vol}(\mathcal{M})} \sum_{i=1}^{k} H_{ii} + 0 = \frac{\operatorname{Tr}(H)}{\operatorname{vol}(\mathcal{M})} = \frac{\Delta_{\mathcal{M}}}{\operatorname{vol}(\mathcal{M})}$$

as desired. For B_t, the same Taylor expansion shows

$$|B_t| \leq \gamma_t \int_{\tilde{B}} e^{-\frac{\|x\|^2}{4t}} O(\|x\|^3)\, dx \xrightarrow{t \to 0} 0$$

For C_t, it shows

$$|C_t| \leq \gamma_t \int_{\tilde{B}} e^{-\frac{\|x\|^2}{8t}} O(\|x\|^5 /t)(1 + O(\|x\|^2))\, dx \xrightarrow{t \to 0} 0$$

where we have used the fact that $\|x\|^2 - g(x) \geq \frac{1}{2} \|x\|^2$ for sufficiently small x.

Therefore

$$\lim_{t \to 0} \frac{1}{t(4\pi t)^{k/2}} \mathbf{L}^t = \lim_{t \to 0} L = \lim_{t \to 0} A_t + B_t + C_t = \frac{\Delta_{\mathcal{M}}}{\mathrm{vol}(\mathcal{M})} \tag{5.19}$$

Step 3: Convergence of $\mathbf{L}_n^t$

We now return to $\mathbf{L}_n^t$, defined above as

$$\mathbf{L}_n^t f(p) = \sum_{j=1}^{n} (f(p) - f(x_i)) e^{\frac{\|p-x_i\|^2}{4t}}$$

We may express $\mathbf{L}_n^t f(p)$ as the sum of n random variables X_i, each corresponding to a data point x_i drawn uniformly from $\mathcal{M}$,

$$\mathbf{L}_n^t f(p) = \sum_{i=1}^{n} X_i \qquad X_i = (f(p) - f(x_i)) e^{\frac{\|p-x_i\|^2}{4t}}$$

with expectation

$$\mathbb{E}[X_i] = \int_{\mathcal{M}} (f(p) - f(y)) e^{\frac{\|p-x\|^2}{4t}}\, d\rho(x)$$

Hoeffding's Inequality states, for $X = \frac{1}{n} \sum_{i=1}^{n} X_i$,

$$\mathbb{P}\left(|X - \mathbb{E}[X]| > \varepsilon\right) \leq 2e^{-\frac{\varepsilon^2 n}{2}}$$

so

$$\mathbb{P}\left(\frac{1}{t(4\pi t)^{k/2}} |X - \mathbb{E}[X]| > \varepsilon\right) \leq 2e^{-\frac{\varepsilon^2 nt(4\pi t)^{k/2}}{2}}$$

Then for fixed $\varepsilon > 0$, if we let $t_n = n^{-\frac{1}{k+2+\alpha}}$, for $\alpha > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{1}{t_n(4\pi t_n)^{k/2}} |X - \mathbb{E}[X]| > \varepsilon\right) = 0$$

From our definitions of $\mathbf{L}_n^t$ and $\mathbf{L}^t$, this is exactly

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{t_n(4\pi t_n)^{k/2}} \cdot \frac{1}{n}\mathbf{L}_n^{t_n} f(p) - \mathbf{L}_n^{t_n} f(p)\right|\right) = 0$$

Thus by Equation 5.19,

$$\lim_{n\to\infty} \frac{1}{t_n(4\pi t_n)^{k/2}} \cdot \frac{1}{n}\mathbf{L}_n^t f(p) = \frac{\Delta_{\mathcal{M}}}{\mathrm{vol}(\mathcal{M})}$$

which completes the proof.

With this result, we conclude our analysis of manifold regularization and the thesis as a whole.

A
Appendix

A.1 Additional Proofs from Chapter 2

A.1.1 Appendix: No Free Lunch

Here, we give a short proof of the No Free Lunch Theorem due to [76]. We work in the setting of binary classification with the 0-1 loss $L(y, f(x)) = 1\{y = f(x)\}$.

Theorem A.1.1 (No Free Lunch)**.** *Let A be a learning algorithm for binary classification on a (potentially infinite) space X. Suppose the training set has size $N < |X|/2$. Then there exists a distribution ρ on $X \times \{0, 1\}$ and a corresponding function $f : X \to \{0, 1\}$ such that:*

1. $L_\rho(f) = 0$
2. $\mathbb{P}(L_\rho(A(S)) \geq 1/8) \geq 1/7$ *for* $S \sim \rho^N$

These two conditions say that for any learning algorithm, there is a distribution ρ that (1) another learner can learn well and (2) it cannot learn well. To be precise, whereas a trivial learner with the single-element hypothesis class $\{f\}$ would incur no loss on ρ, our algorithm incurs a positive loss with constant probability.

Intuitively, when an algorithm $\mathcal{A}$ sees fewer than half the examples in X, there are so many different possible sets of labels for the other examples that at least one will be inconsistent with $\mathcal{A}$.

Proof. Consider a subset $C \subset X$ of size $2N$. Let $f_1, \ldots, f_T$ denote the $T = 2^{2N}$ functions from C to $\{0, 1\}$. For each f_i, let ρ_i be a distribution over $C \times \{0, 1\}$ that is $\frac{1}{2N}$ if f_i maps x to y and 0 otherwise. Mathematically:

$$\rho_i(\{x, y\}) = \tfrac{1}{2N} 1\{y = f_i(x)\}$$

Note that for all i, we have $L_{\rho_i}(f_i) = 0$ by construction.

We now begin the core section of the proof. We will show that an arbitrary $\mathcal{A}$ shown only N out of all $2N$ samples will err with constant probability on average. Precisely, we aim to obtain:

$$\max_{i \in 1 \ldots T} \mathbb{E}_{S \sim \rho_i^N}[L_{\rho_i}(A(S))] \geq 1/4 \tag{A.1}$$

From this result we conclude there exists a ρ and f such that f has 0 loss on data drawn from ρ but $\mathbb{E}_{S \sim \rho^N}[L_\rho(A(S))] \geq 1/4$. It will then be quick to complete the proof by showing how this implies $\mathbb{P}(L_\rho(A(S)) \geq 1/8) \geq 1/7$.

Denote the $K = (2N)^N$ ordered sequences of N examples drawn (with replacement) from C by $S_1, S_2, \ldots, S_K$. Let S_j^i be the result of labeling S_j with function f_i, so $S_j^i = \{(x_1^{(j)}, f_i(x_1^{(j)})), (x_N^{(j)}, f_i(x_N^{(j)}))\}$.

Consider drawing data S from distribution ρ_i. These must all be consistent with f_i, so S may be S_j^i for any $j \in 1, \ldots, K$. Each of these S_j^i are equally likely by symmetry. The expected loss of learning with $\mathcal{A}$ is then the average:

$$\mathbb{E}_{S \sim \rho_i}[L_{\rho_i}(A(S))] = \frac{1}{K} \sum_{j=1}^{K} L_{\rho_i}(A(S_j^i))$$

We now convert our average over sequences to an average over functions:

$$\max_{i \in 1 \ldots T} \frac{1}{K} \sum_{j=1}^{K} L_{\rho_i}(A(S_j^i)) \geq \frac{1}{T \cdot K} \sum_{i=1}^{T} \sum_{j=1}^{K} L_{\rho_i}(A(S_j^i)) \geq \min_{j \in 1 \ldots K} \frac{1}{T} \sum_{i=1}^{T} L_{\rho_i}(A(S_j^i))$$

In words, the worst-case function f_i, averaged over all sequences, incurs at least as much loss as the best-case sequence S_j, averaged over functions.

This switch was useful because we can bound the average performance of $\mathcal{A}$ over distributions ρ_i for fixed j. Fixing j, denote the examples in S_j by

$(x_1, \dots, x_N)$ and denote those not in S_j by $(v_1, \dots, v_L)$. There are at least $L \geq N$ of these examples (not necessarily exactly N due to sampling with replacement).

For every function $g : C \to \{0, 1\}$ and every ρ_i we have:

$$\begin{aligned} L_{\rho_i}(g) &= \frac{1}{2N} \sum_{x \in C} 1\{g(x) \neq f_i(x)\} \geq \frac{1}{2L} \sum_{x \in C} 1\{g(x) \neq f_i(x)\} \\ &\geq \frac{1}{2L} \sum_{x \in C \setminus S_j} 1\{g(x) \neq f_i(x)\} \\ &= \frac{1}{2L} \sum_{l=1}^{L} 1\{g(v_l) \neq f_i(v_l)\} \end{aligned}$$

We can now substitute for g the output $\mathcal{A}(S_j^i)$ of our algorithm and take an average over all T distributions ρ_i.

$$\frac{1}{T} \sum_{i=1}^{T} L_{\rho_i}(\mathcal{A}(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2L} \sum_{l=1}^{L} 1\{\mathcal{A}(S_j^i)(v_l) \neq f_i(v_l)\}$$

Manipulating this with the same average-to-minimum method as above yields a third lower bound, this time involving the examples not in S_j:

$$\frac{1}{T} \sum_{i=1}^{T} L_{\rho_i}(\mathcal{A}(S_j^i)) \geq \frac{1}{2} \min_{p \in 1 \dots L} \frac{1}{T} \sum_{i=1}^{T} 1\{\mathcal{A}(S_j^i)(v_l) \neq f_i(v_l)\}$$

Now fix $l \in 1, \dots, L$. Group the functions $f_1, \dots, f_T$ into the $T/2$ pairs of functions $(f_i, f_{i'})$ that each agree on all examples except v_l. To the learning algorithm $\mathcal{A}$, these pairs of functions look identical because $S_j^i = S_j^{i'}$. They will then make the same predictions on the unseen example v_l: $A(S_j^i)(v_l) = A(S_j^{i'})(v_l)$. However, the two functions in each pair disagree on v_l by construction: $f_i(v_l) \neq f_{i'}(v_l)$.

This implies that the learning algorithm $\mathcal{A}$ will make an incorrect prediction on v_l on half of the distributions ρ:

$$\frac{1}{T} \sum_{i=1}^{T} 1\{\mathcal{A}(S_j^i)(v_l) \neq f_i(v_l)\} = \frac{1}{2}$$

These final steps are the reason that we required $|C| \geq 2N$ in our hypothesis.

Tracing back our steps, we have our core result:

$$\begin{aligned}\max_{i\in 1\ldots T} \mathbb{E}_{S\sim\rho_i^N}[L_{\rho_i}(A(S))] &\geq \min_{j\in 1\ldots K} \frac{1}{T}\sum_{i=1}^{T} L_{\rho_i}(A(S_j^i)) \\ &\geq \frac{1}{2}\min_{p\in 1\ldots L} \frac{1}{T}\sum_{i=1}^{T} 1\{\mathcal{A}(S_j^i)(v_l) \neq f_i(v_l) \\ &\geq \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}\end{aligned}$$

There then must exist a ρ_i for which the expression above holds:

$$\mathbb{E}_{S\sim\rho_i^N}[L_{\rho_i}(A(S))] \geq \frac{1}{4} \tag{A.2}$$

The corresponding function f_i makes no errors on this distribution (Condition 1), but our algorithm A errs a constant fraction of the time.

To finish the proof, we are only left with converting equation A.2 into Condition 2 above. For this, we use Markov's Inequality. For ease of notation, let $Z = L_{\rho_i}(A(S))$. We have that Z is a Bernoulli random variable with $\mathbb{E}[Z] \geq 1/4$. We then have:

$$\begin{aligned}\mathbb{P}(Z \geq a) &= 1 - \mathbb{P}(Z \leq a) \\ &= 1 - \mathbb{P}(1 - Z \geq 1 - a) \\ &\geq 1 - \frac{\mathbb{E}[1-Z]}{1-a} \\ &= \frac{1-\mathbb{E}[Z]}{1-a}\end{aligned}$$

where the third step above holds due to Markov's Inequality. Plugging in $\mathbb{E}[Z] \geq 1/4$ and $a = 1/8$, we have:

$$\mathbb{P}(Z \geq a) \geq \frac{1-1/4}{1-1/8} = \frac{1}{7}$$

Substituting back, we have our final result:

$$\mathbb{P}(L_\rho(\mathcal{A}(S)) \geq 1/8) \geq 1/7$$

for $S \sim \rho^N$. □

A.2 Additional Proofs from Chapter 4

A.2.1 Appendix: Courant-Fischer

The Courant-Fischer Theorem is a consequence of the famous Spectral Theorem.

Theorem A.2.1 (Courant-Fischer)**.** *The k-th smallest eigenvalue λ_k of the self-adjoint matrix $\boldsymbol{A}$ is given by*

$$\lambda_k = \min_{S \subset \mathbb{R}, \dim(S)=k} \max_{x \in S, x \neq 0} \frac{x^T \boldsymbol{A} x}{x^T x} \tag{A.3}$$

where S is a subspace of $\mathbb{R}^n$.

Proof. Since $\mathbf{A}$ is self-adjoint, we may choose an orthonormal basis of eigenvectors $v_1, \dots, v_n$ corresponding to $\lambda_1, \dots, \lambda_n$. Consider a vector $x = \sum_{i=1}^n c_i v_i$. The quadratic form $x^T \mathbf{A} x$ is given by:

$$\begin{aligned} x^T \mathbf{A} x &= \left(\sum_{i=1}^n c_i v_i \right)^T \mathbf{A} \left(\sum_{i=1}^n c_i v_i \right) \\ &= \sum_{i,j=1}^n c_i c_j \lambda_j v_i^T v_j \\ &= \sum_{i=1}^n c_i^2 \lambda_i \end{aligned}$$

We now show that the right side of Equation **??** is upper and lower bounded by λ_k.

For the upper bound, consider $S = \text{span}\,\{v_1, \dots, v_k\}$, so that we can write $x = \sum_{i=1}^k c_i v_i$ for some coefficients c_i. Substituting the expression above for $x^T \mathbf{A} x$ shows

$$\frac{x^T \mathbf{A} x}{x^T x} = \frac{\sum_{i=1}^k c_i^2 \lambda_i}{\sum_{i=1}^k c_i^2} \leq \frac{\sum_{i=1}^k c_i^2 \lambda_k}{\sum_{i=1}^k c_i^2} = \lambda_k$$

It follows that the expression on the right of Equation **??** is upper-bounded by λ_k.

For the lower bound, we need to show that for any choice of S with $\dim(S) = k$, the Rayleigh quotient is at least λ_k. Let

$T = \text{span}\,\{v_k, v_{k+1}, \ldots, v_n\}$. By the same style of argument as above, for $x \in T$, we have

$$\frac{x^T \mathbf{A} x}{x^T x} = \frac{\sum_{i=k}^{n} c_i^2 \lambda_i}{\sum_{i=k}^{k} c_i^2} \leq \frac{\sum_{i=k}^{n} c_i^2 \lambda_k}{\sum_{i=k}^{k} c_i^2} = \lambda_k$$

Also, since $\dim(S) = k$ and $\dim(T) = n - k + 1$, the intersection of S and T has dimension at least 1. Then

$$\max_{x \in S \cap T} \frac{x^T \mathbf{A} x}{x^T x} \leq \max_{x \in T} \frac{x^T \mathbf{A} x}{x^T x} \leq \lambda_k$$

so for any k-dimensional subspace S, there is an x with Rayleigh quotient at most λ_k. This result lower-bounds Equation **??** by λ_k, completing the proof. □

By the same method, we obtain the similar result:

Corollary 4.

$$\lambda_k = \min_{T \subset \mathbb{R}, \dim(T) = n-k-1} \max_{x \in T, x \neq 0} \frac{x^T \boldsymbol{A} x}{x^T x}$$

A.2.2 Appendix: Eigenvalue Bounds (Manifolds)

Theorem A.2.2 (Faber-Krahn Inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary. Let $B \subset \mathbb{R}^n$ be the ball with the same volume as Ω. Denote by $\lambda_2(\Omega)$ the first nonzero eigenvalue of the Laplacian of Ω under Dirichlet boundary conditions ($\partial\Omega = 0$). Then*

$$\lambda_2(\Omega) \geq \lambda_2(B)$$

with equality if and only if $\Omega = B$.

The following proof is due to [56].

Proof of Faber-Krahn. Denote by f the eigenfunction corresponding to $\lambda_2(\Omega)$. We will construct a radial function g on the ball B that resembles f. Define $g : B \to \mathbb{R}^+$ to be the radial function such that

$$\text{vol}(f \geq t) = \text{vol}(g \geq t)$$

That is,

$$g(x) = \sup\left\{t \geq 0 : \text{vol}(f \geq t) \geq \text{vol}(B_{\|x\|})\right\}$$

We have constructed g in this manner so that integrating over t gives:

$$\int_\Omega f^2\,dV = \int_0^\infty \mathrm{vol}(f^2 \geq t)\,dV = \int_0^\infty \mathrm{vol}(g^2 \geq t)\,dV = \int_B g^2\,dV$$

Using the Rayleigh quotient characterization of the eigenvalue λ_2, we have

$$\lambda_2(\Omega) = \frac{\int_\Omega |\nabla f|^2}{\int_\Omega f^2} \qquad \text{and} \qquad \lambda_2(B) = \frac{\int_B |\nabla g|^2}{\int_B g^2}$$

We have shown that the denominators are equal, so it remains to be shown that $\int_\Omega |\nabla f|^2 \geq \int_\Omega |\nabla g|^2$.

Consider the area of a level set $\{g = t\}$. Since g is radial, it is constant on its own level sets:

$$\mathrm{Area}\{g = t\} = \int_{\{g=t\}} dS = \sqrt{\int_{\{g=t\}} |\nabla g|\,dS \int_{\{g=t\}} \frac{1}{|\nabla g|}\,dS}$$

For f, by Cauchy-Schwartz:

$$\mathrm{Area}\{f = t\} = \int_{\{f=t\}} dS \leq \sqrt{\int_{\{f=t\}} |\nabla f|\,dS \int_{\{f=t\}} \frac{1}{|\nabla f|}\,dS}$$

The key step of the proof is to use the isoperimetric inequality, which states that the ball is the surface with maximal ratio of volume to surface area.

$$\begin{aligned}\sqrt{\int_{\{f=t\}} |\nabla f|\,dS \int_{\{f=t\}} \frac{1}{|\nabla f|}\,dS} &\geq \mathrm{Area}\{f = t\} \geq \mathrm{Area}\{g = t\} \\ &= \sqrt{\int_{\{g=t\}} |\nabla g|\,dS \int_{\{g=t\}} \frac{1}{|\nabla g|}\,dS} \qquad \text{(A.4)}\end{aligned}$$

Next, the co-area formula states

$$\mathrm{vol}(\Omega') = \int_{\Omega'} dV = \int_{-\infty}^\infty \frac{1}{|\nabla f|} \mathrm{Area}(f^{-1}(t))dt$$

which applied to f on Ω and g on B gives:

$$\int_{\{f=t\}} \frac{1}{|\nabla f|}\,dS = -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vol}(f \geq t) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vol}(g \geq t) = \int_{\{g=t\}} \frac{1}{|\nabla g|}\,dS \qquad \text{(A.5)}$$

where the middle equality holds because $\mathrm{vol}(f \geq t) = \mathrm{vol}(g \geq t)$.

From Equations A.4 and A.5, we see

$$\int_{\{f=t\}} |\nabla f|\, dS \geq \int_{\{g=t\}} |\nabla g|\, dS$$

and so

$$\int_\Omega |\nabla f|^2 = \int_0^\infty \left(\int_{\{f=t\}} |\nabla f|\, dS \right) dt \geq \int_0^\infty \left(\int_{\{g=t\}} |\nabla g|\, dS \right) dt = \int_\Omega |\nabla g|^2$$

This result completes the proof. □

A.2.3 Appendix: Eigenvalue Bounds (Graphs)

The following theorem was proven by Miroslav Fiedler in 1973 [32] and is the origin of the term "Fielder value".

Theorem A.2.3 (Fielder)**.**

$$\lambda_2 \leq \frac{n}{n-1} \min_{v \in V} d_v \qquad \textit{and} \qquad \lambda_n \geq \frac{n}{n-1} \max_{v \in V} d_v \tag{A.6}$$

Proof. Define the matrix M by

$$M = \mathbf{L} - \lambda_2(I - J/n)$$

Note that $M\mathbf{1} = 0$ for the constant vector $\mathbf{1}$ because $(I - J/n)\mathbf{1} = 0$.

Any vector y may be decomposed into its orthogonal components $y = c_1\mathbf{1} + c_2 x$, where x is a unit-length vector orthogonal to $\mathbf{1}$. Then we have

$$y^T M y = c_2^2 x^T M x = c_2^2 (x^T \mathbf{L} x - \lambda_2)$$

Since $\lambda_2 = \min_{x \perp \mathbf{1}, \|x\|^2 = 1} x^T \mathbf{L} x$, the quantity above is always positive, so that M is positive semidefinite.

Let M_{ii} denote the i-th diagonal element of M. Note that $M_{ii} \geq 0$ (as it equals $e_i^T M e_i$). We then have

$$\min_i M_{ii} = \min_i L_{ii} - \lambda_2(1 - 1/n) \geq 0$$

and rearranging gives A.6. □

A bound on λ_n was proven by Anderson and Morley in 1985 [3].

Theorem A.2.4 (Anderson and Morley)**.**

$$\lambda_n \leq \max_{(i,j)\in E}(d_i + d_j)$$

This bound was strengthened by Merris [64], who also provided a simple proof based on Gershgorin's circle theorem.

Theorem A.2.5 (Merris)**.** *Let $m(i)$ be the average of the degrees of vertices adjacent to vertex i. That is, $m(i) = \frac{1}{|N(i)|}\sum_{j\in N(i)} d_j$ where $N(i)$ denotes the neighbors of i. Then*

$$\lambda_n \leq \max_{i\in V}(d_i + m(i)) \tag{A.7}$$

Lemma A.2.1 (Gershgorin's circle theorem)**.** *Let M be an $n \times n$ matrix with entries m_{ij}. Let $r_i = \sum_{j\neq i}|m_{ij}|$ be the sum of the non-diagonal elements of the i-th row of M. Let $D_i = D(m_{ii}, r_i) \subset \mathbb{C}$ be the closed disk in the complex plane with radius r_i and center m_{ii}. Then every eigenvalue of M is contained in some D_i.*

Proof of Lemma. Let λ be an eigenvalue of M with corresponding eigenvector v. Without loss of generality, let the component v_i of v with largest magnitude be 1. We have

$$(Mv)_i = (\lambda v)_i = \lambda$$

and

$$(Mv)_i = \sum_j m_{ij}v_j = \sum_{j\neq i} m_{ij}v_j + m_{ii}$$

so then

$$|lam - m_{ii}| = \left|\sum_{j\neq i} m_{ij}v_j\right| \leq \sum_{j\neq i}|m_{ij}||v_j| \leq \sum_{j\neq i}|m_{ij}| = r_i$$

showing that $\lambda \in D_i$. □

Proof of Merris' Bound. Consider $\overline{L} = D^{-1}\mathbf{L}D$, where D is the diagonal matrix of degrees of vertices.

$$\overline{L}_{ij} = \begin{cases} d_i & i = j \\ -d_j/d_i & (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Applying Gershgorin's circle theorem gives that every eigenvalue λ of $\overline{L}$ is bounded by

$$\max_i \overline{L}_{ii} + r_i = \max_i \overline{L}_{ii} + \sum_{j \in N(i)} |-d_j/d_i| = \max_i (d_i + \frac{1}{N(i)} \sum_{j \in N(i)} d_j) = \max_i (d_i + m(i))$$

Since $\mathbf{L}$ is similar to $D^{-1}\mathbf{L}D$, they share the same eigenvalues, and A.7 holds. □

A simple bound relates λ_2 on a graph to λ_2 on a subset of the vertices.

Theorem A.2.6. *For a subset $S \subset V$ of the vertices of G, let $G \setminus S$ denote the graph with all vertices in S and edges connecting to S removed. Then*

$$\lambda_2(G) \leq \lambda_2(G \setminus S) + |S|$$

Proof. Let v be an eigenvector of the Laplacian of $G \setminus S$ corresponding to the eigenvalue $\lambda_2(G \setminus S)$. Consider v as a vector on all of G by adding 0s in the entries corresponding to 0. By the Rayleigh characterization of λ_2,

$$\lambda_2 \leq \sum_{(i,j) \in E(G)} (v_i - v_j)^2$$

Each of these edges has $0, 1$, or 2 vertices in S, so

$$\lambda_2 \leq \sum_{(i,j) \in E(G \setminus S)} (v_i - v_j)^2 + \sum_{i \in S} \sum_{j \in N(i)} v_j^2 + 0 \leq \lambda_2(G \setminus S) + |S|$$

□

A.2.4 Appendix: Cauchy's Interlacing Theorem

Cauchy's Interlacing Theorem is a satisfying result relating the eigenvalues of a matrix to those of a principal submatrix of dimension $(n-1)$ (i.e. a submatrix obtained by deleting the same row and column). As one might intuitively expect, these set of eigenvalues cannot differ greatly.

We prove two versions of this result, the second of which is sometimes called Weyl's Theorem or Weyl's Perturbation Inequality.

Theorem A.2.7 (Cauchy's Interlacing Theorem)**.** *Let A be a self-adjoint $n \times n$ matrix. Let B be a principal submatrix of A of dimension $n-1$. Denote the eigenvalues of A and B by $\alpha_1 \leq \cdots \leq \alpha_n$ and $\beta_1 \leq \cdots \leq \beta_n$, respectively. Then*

$$\alpha_1 \leq \beta_1 \leq \alpha_2 \leq \cdots \leq \alpha_{n-1} \leq \beta_n \leq \alpha_n$$

Proof. Without loss of generality, let the first row and column of A be deleted. By the Courant-Fischer Theorem applied to A,

$$\alpha_{k+1} = \max_{S \subset \mathbb{R}, \dim(S)=n-k} \min_{x \in S, x \neq 0} \frac{x^T A x}{x^T x}$$

and by the Courant-Fischer Theorem applied to B,

$$\beta_k = \max_{S \subset \mathbb{R}^{n-1}, \dim(S)=n-k-1} \min_{x \in S, x \neq 0} \frac{x^T B x}{x^T x} = \max_{S \subset \mathbb{R}^{n-1}, \dim(S)=n-k-1} \min_{x \in S, x \neq 0} \frac{(0 \;\; x)^T A (0 \;\; x)}{x^T x}$$

where $(0 \;\; x)$ is the n-dimensional vector with 0 in its first component and the entries of x in its $(n-1)$ other components. Comparing these expressions, we see $\alpha_{k+1} \geq \beta_k$ because the expression for β_k is the same as that for α_k, but taken over a smaller space. The other direction ($\alpha_k \leq \beta_k$) is obtained by the same method applied to α_k and β_k. □

Corollary 5. *Let B be a principal submatrix of A of dimension r. Then*

$$\alpha_i \leq \beta_i \leq \alpha_{i+n-r}$$

Proof. Apply Cauchy's Interlacing Theorem r times. □

An application of these ideas is that removing an edge from a graph decreases its eigenvalues. The proof here, due to [36], uses heavy machinery from complex analysis.

Theorem A.2.8 (Edges Increase Eigenvalues)**.** *Let G be a non-complete graph and (i,j) an edge not in E. Denote by G' the graph G with edge (i,j) added. Then the eigenvalues of G' interlace those of G:*

$$0 = \lambda_1(G) = \lambda_1(G') \leq \lambda_2(G) \leq \lambda_2(G') \leq \lambda_3(G) \leq \cdots \leq \lambda_n(G) \leq \lambda_n(G')$$

Proof. Let L and L' be the Laplacians of G and G', respectively. Let z be the vector that is 1 in the entry corresponding to vertex i, -1 in the entry corresponding to vertex j, and 0 elsewhere. Then $L' = L - zz^T$.

For a real number t, consider the quantity $tI - L'$. We have

$$tI - L' = tI - L - zz^T = (tI - L)(I - (tI - L)^{-1}zz^T)$$

Taking determinants gives:

$$\det(tI - L') = \det(tI - L)\det(I - (tI - L)^{-1}zz^T)$$

The determinant has the property that $\det(I - CD) = \det(I - DC)$, so

$$\det(I - (tI - L)^{-1}zz^T) = 1 - z^T(tI - L)^{-1}z$$

and

$$\frac{\det(tI - L')}{\det(tI - L)} = 1 - z^T(tI - L)^{-1}z$$

Denote this expression as a function of t by $\psi(t)$.

We now prove a lemma about rational functions of this form.

Lemma A.2.2. *Let ψ be a rational function of the form*
$\psi(t) = z^T(tI - L)^{-1}z$ for a real self-adjoint matrix L. Then

1. *ψ has simple zeros and poles*

2. *$\psi' < 0$ where it is defined.*

3. *Consecutive poles of ψ are separated by no more than 1 zero of ψ.*

Proof of Lemma. Write

$$\psi(t) = \sum_{\lambda \in \mathrm{eval}(L)} \frac{z^T v_\lambda z}{t - \lambda}$$

where $\mathrm{eval}(L)$ denotes the set of eigenvalues of L with corresponding eigenvectors v_λ. Note that the poles of this expression are simple.

Differentiating gives

$$\psi'(t) = -\sum_{\lambda \in \mathrm{eval}(L)} \frac{z^T v_\lambda z}{t - \lambda}^2 = -z^T(tI - L)^{-2}z$$

which is negative as $z^T(tI-L)^{-2}z = \left\|(tI-L)^{-1}z\right\|^2$. Then each zero of ψ is simple.

Now consider consecutive poles a and b of ψ. As they are simple and $\psi' < 0$, ψ is strictly decreasing on $[a, b]$. Since t is positive near a in this interval and negative near b in this interval, it follows that ψ has exactly one zero in $[a, b]$. This result completes the lemma. □

We now complete the main proof. Applying the lemma with $\psi(t)$ defined as above, we see that ψ has simple zeros and poles, with consecutive poles separated by a single zero. Its poles are the zeros of $\det(tI - L)$ and its zeros are the zeros of $\det(tI - L')$. In other words, its poles are the eigenvalues of L and its zeros are the eigenvalues of L'. It follows from the lemma that the n zeros and poles of ψ interlace.

It remains to be shown that this interlacing begins with an eigenvalues of L (and not L'), but this is clear because the trace of L' (the sum of the eigenvalues) is 2 greater than the trace of L. □

A.2.5 Appendix: Cheeger's Inequality

Cheeger's Inequality relates the conductance of a graph or manifold to its second eigenvalue λ_2.

Theorem A.2.9 (Cheeger's Inequality for Graphs)**.** *For an unweighted d-regular graph,*

$$h(G) \leq \sqrt{2d\lambda_2}$$

Theorem A.2.10 (Cheeger's Inequality for Manifolds)**.** *For a closed manifold $\mathcal{M}$,*

$$h(\mathcal{M}) \leq \sqrt{2\lambda_2}$$

The following proofs are due to [83].

Proof (Graphs). The proof is based on the Rayleigh characterization of λ_2. For ease of notation, alongside the Rayleigh quotient $R(f)$, define the L^1 Rayleigh quotient $R^1(f)$ as

$$R^1(f) = \frac{\sum_{(i,j)\in E} |f(i) - f(j)|}{\sum_{(i,j)\in E} |f(i)|}$$

As an aside, note that we used the L^1 Rayleigh quotient above (without defining it) to measure the boundary of subsets.

The proof proceeds in three lemmas. The outline is as follows:

1. First, we show there exists a nonnegative function $\hat{f}$ supported on at most half the vertices of G such that $R(\hat{f}) \leq \lambda_2$.
2. Second, we consider the elementwise square of $\hat{f}$, denoted g. We show
$$R^1(g) \leq \sqrt{2dR(\hat{f})}$$
3. Third, we show there exists a real $t \geq 0$ such that the set $S = \{i : g(i) > t\}$ has
$$h_G(S) \leq R^1(g)$$
Then $h(G) \leq h_G(S) \leq R^1(g) \leq \sqrt{2dR(\hat{f})} \leq \sqrt{2d\lambda_2}$.

Lemma A.2.3 (G1)**.** *Let f be a vector orthogonal to the constant vector. Then there exists a vector $\hat{f}$ with nonnegative entries such that:*

1. $|\{i : \hat{f}(i) > 0\}| \leq \frac{1}{2}|V|$

2. $R(\hat{f}) \leq R(f)$

Proof. Denote by m the median of the entires of f. Let $\overline{f} = f - m\mathbf{1}$, where $\mathbf{1}$ is the constant vector of 1s. We have

$$\langle \overline{f}, \Delta \overline{f} \rangle \langle f - m\mathbf{1}, \Delta(f - m\mathbf{1}) \rangle = 0 + \langle f, \Delta f \rangle$$

and

$$\langle \overline{f}, \overline{f} \rangle = \langle f - m\mathbf{1}, f - m\mathbf{1} \rangle = \langle f, f \rangle + \langle m\mathbf{1}, m\mathbf{1} \rangle \geq \langle f, f \rangle$$

because $f \perp \mathbf{1}$ and $\Delta \mathbf{1} = 0$. Then

$$R(\overline{f}) = \frac{\langle \overline{f}, \Delta \overline{f} \rangle \langle}{\langle \overline{f}, \overline{f} \rangle} \leq \frac{\langle f, \Delta f \rangle \langle}{\langle f, f \rangle} = R(f) = \lambda_2$$

Now split f into two vectors consisting of its positive and negative components, $f = f^+ - f^-$. That is, $f_i^+ = \max(0, \overline{f}_i)$ and $f_i^- = \max(0, -\overline{f}_i)$.

Let $\hat{f}$ be the vector in $\{f^+, f^-\}$ with smaller Rayleigh quotient.

$$\hat{f} = \begin{cases} f^+ & R(f^+) < R(f^-) \\ f^- & \textit{otherwise} \end{cases}$$

Since both f^+ and f^- have at most $|V|/2$ nonzero entries, $\hat{f}$ is supported on at most half the vertices of G. It remains to bound $\min(R(f^+), R(f^-))$.

Using the fact that for $a_1, b_1, a_2, b_2 > 0$,

$$\min\left(\frac{a_1}{b_1}, \frac{a_2}{b_2}\right) \leq \frac{a_1 + a_2}{b_1 + b_2}$$

we obtain

$$\begin{aligned} \min(R(f^+), R(f^-)) &= \min\left(\frac{\langle f^+, \Delta f^+\rangle}{\langle f^+, f^+\rangle}, \frac{\langle f^-, \Delta f^-\rangle}{\langle f^-, f^-\rangle}\right) \\ &\leq \frac{\langle f^+, \Delta f^+\rangle + \langle f^-, \Delta f^-\rangle}{\langle f^+, f^+\rangle + \langle f^-, f^-\rangle} \end{aligned}$$

Since f^+ and f^- have disjoint support, $\langle f^+, f^-\rangle = 0$ and

$$\langle f^+, f^+\rangle + \langle f^-, f^-\rangle = \langle f^+ - f^-, f^+ - f^-\rangle = \langle \overline{f}, \overline{f}\rangle$$

Also, by the triangle inequality,

$$\langle f^+, \Delta f^+\rangle + \langle f^-, \Delta f^-\rangle \leq \langle \overline{f}, \Delta \overline{f}\rangle$$

As a result,

$$\begin{aligned} \min(R(f^+), R(f^-)) &\leq \frac{\langle \overline{f}, \Delta \overline{f}\rangle}{\langle f^+, f^+\rangle + \langle f^-, f^-\rangle} \\ &= R(\overline{f}) \leq R(f) = \lambda_2 \end{aligned}$$

which completes the proof of the lemma. □

Lemma A.2.4 (G2). *For a vector f, if g is defined by $g_i = f_i^2$, then $R^1(g) \leq \sqrt{2dR(f)}$.*

Proof. This lemma is the Cauchy-Schwartz inequality in disguise. Applying

Cauchy-Schwartz to the numerator of $R^1(g)$ gives

$$\begin{aligned}\sum_{(i,j)\in E}|g(i)-g(j)| &= \sum_{(i,j)\in E}|f^2(i)-f^2(j)| \\ &= \sum_{(i,j)\in E}|f(i)-f(j)|(f(i)+f(j)) \\ &\leq \sqrt{\sum_{(i,j)\in E}(f(i)-f(j))^2}\sqrt{\sum_{(i,j)\in E}(f(i)+f(j))^2} \qquad \text{(CS)} \\ &= \sqrt{\sum_{(i,j)\in E}R(f)\sum_i f(i)^2}\sqrt{\sum_{(i,j)\in E}(f(i)+f(j))^2} \qquad \text{(def of } R(f)\text{)} \\ &\leq \sqrt{\sum_{(i,j)\in E}R(f)\sum_i f(i)^2}\sqrt{\sum_{(i,j)\in E}2f(i)^2+2f(j)^2} \\ &= \sqrt{R(f)\sum_i f(i)^2}\sqrt{2d\sum_i f(i)^2} \\ &= \sqrt{2dR(f)}\sum_i f(i)^2 = \sqrt{2dR(f)}\sum_i g(i)\end{aligned}$$

Therefore

$$R^1(g) = \frac{\sum_{(i,j)\in E}|g(i)-g(j)|}{\sum_i g(i)} \leq \sqrt{2dR(f)}$$

□

Lemma A.2.5 (G3)**.** *For every nonnegative vector g, there is a real $t > 0$ such that*

$$\frac{|\partial\{i : g(i) > t\}|}{|\{i : g(i) > t\}|} \leq R^1(g)$$

Proof. Let $S_t = \{i : g(i) > t\}$. These S_t are sometimes called *Sweep sets.* For each edge (i,j), let $\mathbf{1}_{ij}^t$ denote the indicator that $(i,j) \in S_t$.

First, we relate $|\partial S_t|$ to $R^1(g)$. The numerator of $R_1(g)$ may be expressed as

$$\sum_{(i,j)\in E}|g(i)-g(j)| = \sum_{(i,j)\in E}\int_0^\infty \mathbf{1}_{ij}^t dt$$

The size of the boundary of S_t is $|\partial S_t| = \sum_{(i,j)\in E} \mathbf{1}_{ij}^t$, so

$$\int_0^\infty |\partial S_t| = \sum_{(i,j)\in E} |g(i) - g(j)|$$

Also note that

$$\int_0^\infty |S_t| = \sum_i g(i)$$

Putting these together, we have

$$\begin{aligned} R^1(g) &= \frac{\sum_{(i,j)\in E} |g(i) - g(j)|}{\sum_i g(i)} \\ &\leq \frac{\int_0^\infty |\partial S_t|}{\int_0^\infty |S_t|} \end{aligned}$$

Letting t^* be the minimizer of $|\partial S_t|/|S_t|$, we have

$$\begin{aligned} R^1(g) &\leq \frac{\int_0^\infty |\partial S_{t^*}|/|S_{t^*}||S_t|}{\int_0^\infty |S_t|} \\ &= |\partial S_{t^*}|/|S_{t^*}| \end{aligned}$$

Therefore t^* satisfies the statement of the lemma. □

We may now complete the proof of Cheeger's Inequality. Let f be the eigenfunction corresponding to λ_2. By Lemma G1, we obtain a corresponding nonnegative function $\hat{f}$, supported on at most half the vertices of G, such that $R(\hat{f}) \leq R(f) = \lambda_2$. By Lemma G2, with g denoting the elementwise square of $\hat{f}$, we obtain

$$R^1(g) \leq \sqrt{2dR(\hat{f})}$$

Apply Lemma G3 and denote the resulting set by $S = \{i : g(i) > t\}$. By the lemma and the fact that S contains at most half the vertices of G, we have

$$h_G(S) = \frac{|\partial S|}{|S|} \leq R^1(g) \leq \sqrt{2d\hat{f}} \leq \sqrt{2d\lambda_2}$$

□

The proof of the manifold case follows a nearly identical structure.

Proof (Manifolds). Similarly to the proof above, define the L^1 Rayleigh quotient $R^1(f)$ as

$$R^1(f) = \frac{\int_{\mathcal{M}} \|\nabla f\| \, dV}{\int_{\mathcal{M}} \|f\| \, dV}$$

The proof proceeds in three lemmas.

1. First, we show there exists a nonnegative function $\hat{f}$ with supported on a set of volume at most $\frac{1}{2}\text{vol}(\mathcal{M})$ such that $R(\hat{f}) \leq \lambda_2$.
2. Second, we show that

   $$R^1(f^2) \leq \sqrt{2R(\hat{f})}$$

3. Third, we show there exists a real $t \geq 0$ such that the set $S = \{x : f^2(x) > t\}$ has

   $$h_{\mathcal{M}}(S) \leq R^1(f^2)$$

   Then we have

   $$h(\mathcal{M}) \leq h_{\mathcal{M}}(S) \leq R^1(f^2) \leq \sqrt{2R(\hat{f})} \leq \sqrt{2\lambda_2}$$

Lemma A.2.6 (M1 (Manifolds))**.** *Let f be a function with $\int_{\mathcal{M}} f = 0$. Then there exists a function $\hat{f} \geq 0$ such that:*

1. *vol*$(\{x : \hat{f}(x) > 0\}) \leq \frac{1}{2}$*vol*$(\mathcal{M})$
2. $R(\hat{f}) \leq R(f)$

Proof. Let m be a median of f, which is to say the smallest m such that vol$(\{x : f(x) < m\}) \geq 1/2$. Let $\overline{f}(x) = f(x) - m$. The numerators of $R(\overline{f})$ and $R(f)$ are the same

$$\langle \overline{f}, \Delta \overline{f} \rangle \langle f - m, \Delta(f - m) \rangle = 0 + \langle f, \Delta f \rangle$$

since the Laplacian of a constant is 0. The denominator of $R(\overline{f})$ is larger

$$\langle \overline{f}, \overline{f} \rangle = \langle f - m, f - m \rangle = \langle f, f \rangle + \langle m, m \rangle \geq \langle f, f \rangle$$

because f is orthogonal to a constant (i.e. it integrates to 0). Note that whereas in the proof above, these inner products referred to matrix products, here they refer to integration over $\mathcal{M}$.

We then have

$$R(\overline{f}) = \frac{\langle \overline{f}, \Delta \overline{f} \rangle \langle}{\langle \overline{f}, \overline{f} \rangle} \leq \frac{\langle f, \Delta f \rangle \langle}{\langle f, f \rangle} = R(f) = \lambda_2$$

Now let $f_i^+ = \max(0, \overline{f}_i)$ and $f_i^- = \max(0, -\overline{f}_i)$. Define $\hat{f}$ be the function in $\{f^+, f^-\}$ with smaller Rayleigh quotient. Note that $\hat{f}$ is supported on a region with volume at most half of that of $\mathcal{M}$.

The remainder of the proof is exactly the same as the proof for graphs above.

$$\begin{aligned} \min(R(f^+), R(f^-)) &= \min\left(\frac{\langle f^+, \Delta f^+ \rangle}{\langle f^+, f^+ \rangle}, \frac{\langle f^-, \Delta f^- \rangle}{\langle f^-, f^- \rangle} \right) \\ &\leq \frac{\langle f^+, \Delta f^+ \rangle + \langle f^-, \Delta f^- \rangle}{\langle f^+, f^+ \rangle + \langle f^-, f^- \rangle} \end{aligned}$$

Since f^+ and f^- have disjoint support, $\langle f^+, f^- \rangle = 0$ and

$$\langle f^+, f^+ \rangle + \langle f^-, f^- \rangle = \langle f^+ - f^-, f^+ - f^- \rangle = \langle \overline{f}, \overline{f} \rangle$$

Also, by the triangle inequality,

$$\langle f^+, \Delta f^+ \rangle + \langle f^-, \Delta f^- \rangle \leq \langle \overline{f}, \Delta \overline{f} \rangle$$

As a result,

$$\begin{aligned} \min(R(f^+), R(f^-)) &\leq \frac{\langle \overline{f}, \Delta \overline{f} \rangle}{\langle f^+, f^+ \rangle + \langle f^-, f^- \rangle} \\ &= R(\overline{f}) \leq R(f) = \lambda_2 \end{aligned}$$

which completes the proof of the lemma. □

Lemma A.2.7 (M2)**.** *For nonnegative f,*

$$R^1(f^2) \leq \sqrt{2R(f)}$$

Proof. We apply the chain rule and the Cauchy-Schwartz inequality:

$$\begin{aligned}
\int_{\mathcal{M}} \left\|\nabla(f^2)\right\| \, dV &= \int_{\mathcal{M}} 2|f| \, \|\nabla f\| \, dV && \text{(Chain Rule)} \\
&\leq \sqrt{\int_{\mathcal{M}} 4f^2 \, dV} \sqrt{\int_{\mathcal{M}} \|\nabla f\|^2 \, dV} && \text{(CS)} \\
&= 2 \int_{\mathcal{M}} f^2 \, dV \cdot \sqrt{R(f)}
\end{aligned}$$

Therefore

$$R^1(f^2) \leq \sqrt{2R(f)}$$

□

Lemma A.2.8 (M3)**.** *For every nonnegative function g, there is a real $t > 0$ such that*

$$\frac{\textit{area}(\partial\{x : g(x) > t\})|}{\textit{vol}(\{x : g(x) > t\})} \leq R^1(g)$$

Proof. Let $S_t = \{x : g(x) > t\}$. Consider the numerator and denominator of $R^1(g)$.

For the numerator, the coarea formula (4.10) states

$$\int_{\mathcal{M}} \|\nabla g\| \, dV = \int_0^\infty \text{area}(\partial S_t) \, dt$$

For the denominator, observe that

$$\int_{\mathcal{M}} \|g\| \, dV = \int_0^\infty \text{vol}(S_t) \, dt$$

Putting these together, we have

$$R^1(g) = \frac{\int_0^\infty \text{area}(\partial S_t) \, dt}{\int_0^\infty \text{vol}(S_t) \, dt}$$

Letting t^* be the minimizer of $\text{area}(\partial S_t)/\text{vol}(S_t)$, we have

$$\begin{aligned}
R^1(g) &\leq \frac{\int_0^\infty \text{area}(\partial S_{t^*})/\text{vol}(S_{t^*})\text{vol}(S_t)}{\int_0^\infty \text{vol}(S_t)} \\
&= \text{area}(\partial S_{t^*})/\text{vol}(S_{t^*})
\end{aligned}$$

Therefore t^* satisfies the statement of the lemma. □

To complete the proof of Cheeger's Inequality on manifolds, let f be the eigenfunction corresponding to λ_2. By Lemma M1 (Manifolds), we obtain a function $\hat{f}$ supported on a set with volume at most half that of $\mathcal{M}$, such that $R(\hat{f}) \leq R(f) = \lambda_2$. By Lemma M2, we obtain $R^1(f^2) \leq \sqrt{2R(\hat{f})}$. Apply Lemma M3 and denote the result by $S = \{x : g(x) > t\}$. By the lemma and the fact that $\text{vol}(S) \leq \frac{1}{2}\text{vol}(\mathcal{M})$,

$$h_G(S) = \frac{|\partial S|}{|S|} \leq R^1(g) \leq \sqrt{2\hat{f}} \leq \sqrt{2\lambda_2}$$

□

Upon proving Cheeger's inequality, we have a few remarks. First, Cheeger's inequality is tight; the path graph, which we saw above, has

$$h(G) = 1/\lceil (n-1)/2 \rceil \qquad \text{and} \qquad \lambda_2 \approx \frac{\pi^2}{2(n-1)^2}$$

Second, the proof of Cheeger's inequality for graphs immediately yields an algorithm for finding a subset of vertices with $h_G(S) \leq \sqrt{2\lambda_2}$. Such a set is called a *sparse cut* of G.

Algorithm 1: Finding a sparse cut from f_2

```
Input: The $2^{\text{nd}}$ eigenfunction $f_2$
Result: A sparse cut $S \subset V$
$f \leftarrow D^{-1/2} f_2$
Sort the vertices so $f(v_1) \leq \cdots \leq f(v_n)$
Initialize $i \leftarrow 0$,   $S \leftarrow \emptyset$,   $S^* \leftarrow \{v_1\}$
while $i < n$ do
    $i = i + 1$
    $S = S \cup \{v_i\}$
    if $h_G(S) \leq h_G(S^*)$ then
        $S^* \leftarrow S$
    end
end
return $S^*$
```

Theorem A.2.11. *Let $u(x,t)$ be a solution to the homogeneous heat equation. Then $\phi(t) = \|u(\cdot,t)\|_{L^2}$ is a nonincreasing function of t.*

Proof.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|u(\cdot,t)\|_{L^2} &= 2\int_{\mathcal{M}} \partial_t u(x,t)u(x,t)\,d\mu(x)\\ &= -2\int_{\mathcal{M}} \Delta u(x,t)u(x,t)\,d\mu(x)\\ &= -2\|\nabla u(\cdot,t)\|^2\end{aligned}$$

Since the derivative of $\phi(t)$ is always negative, it is a nonincreasing function of t. □

Theorem A.2.12. *A solution to the homogeneous heat equation is unique.*

Proof. Suppose u_1 and u_2 solve the homogeneous heat equation. Then $u = u_1 - u_2$ solves

$$\begin{aligned}Lu(x,t) &= 0\\ u(x,0) &= 0\end{aligned}$$

By the theorem above, the function $t \mapsto \int_{\mathcal{M}} u(x,t)^2\,dx$ is a nonincreasing function of t. Since $u(x,0) = 0$, we must have $u(x,t) = 0$. Therefore $u_1 = u_2$. □

Theorem A.2.13 (Sturm-Liouville decomposition)**.** *Denote the eigenvalues and eigenfunctions of the Laplacian Δ by $\lambda_1 \le \lambda_2 \le \cdots$ and $\phi_1, \phi_2, \dots,$ respectively. Then*

$$p(x,y,t) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(x)\phi_i(y)$$

The following proof is adopted from [18].

Proof. By the spectral theorem, as $e^{-\Delta}$ is a compact self-adjoint operator, it has eigenvalues

$$\beta_1 \ge \beta_2 \ge \cdots$$

with corresponding eigenfunctions $\phi_1, \dots, \phi_n$.

Let $\lambda_i = -\ln \beta_i$. We aim to show these λ_i are the eigenvalues of Δ. By the properties of the heat operator,

$$e^{-t\Delta}\phi_k = \left(e^{-\Delta}\right)^t \phi_k = \beta_k^t \phi_k = e^{-t\lambda_k}\phi_k$$

As $e^{-t\Delta}\phi_k$ solves the heat equation, we have

$$\begin{aligned} 0 &= L(e^{-t\Delta}\phi_k) = L(e^{-t\lambda}\phi_k) \\ &= \Delta e^{-t\lambda}\phi_k + \partial_t e^{-t\lambda}\phi_k \\ &= e^{-t\lambda}(\Delta\phi_k - \lambda_k\phi_k) \end{aligned}$$

so $\Delta\phi_k = \lambda_k\phi_k$, and λ_k is an eigenvalue of Δ corresponding to eigenfunction ϕ_k.

Note that by the definition of the heat propagator,

$$\langle p(x,\cdot,t), \phi_k\rangle\phi_k(y) = \int_{\mathcal{M}} p(x,y,t)\phi_k(y)\,d\mu(y) = e^{-t\Delta}\phi_k(x) = e^{-t\lambda_k}\phi_k(x)$$

Finally, since the ϕ_i form a basis for $L^2(\mathcal{M})$, we can write p as

$$p(x,y,t) = \sum_{k=0}^{\infty}\langle p(x,\cdot,t), \phi_k\rangle\phi_k(y) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(x)\phi_i(y)$$

□

A.3 Additional Proofs from Chapter 5

A.3.1 Appendix: Integral Operators

Lemma A.3.1. *The functions $\sqrt{\lambda_i}e_i$ form an orthonormal basis for $\mathcal{H}_K$.*

Proof. First, we show the collection $\{\sqrt{\lambda_i}e_i\}$ are orthonormal in $\mathcal{H}_K$. Observe that

$$\langle K_x, e_i\rangle_\rho = \int f(y)K(x,y)\,d\rho(y) = (I_K e_i)(x) = \lambda_i e_i(x)$$

so $K_x = K(x,\cdot) = \sum_{i=1}^{\infty}\lambda_i e_i(x)e_i$. Then by the reproducing property,

$$e_j(x) = \langle K_x, e_j\rangle_K = \sum_{i=1}^{\infty}\lambda_i e_i(x)\langle e_i, e_j\rangle_K$$

which implies

$$\langle e_i, e_j \rangle_K = \begin{cases} 0 & i \neq j \\ 1/\lambda_i & i = j \end{cases}$$

Therefore the rescaled vectors $\{\sqrt{\lambda_i} e_i\}$ are orthonormal in $\mathcal{H}_K$.

Second, we show that $\{e_i\}$ spans $\mathcal{H}_K$. Let $f \in \mathcal{H}_K$ be orthogonal to e_i for all i. Then by the reproducing property:

$$\begin{aligned} f(x) = \langle f, K_x \rangle_K &= \left\langle f, \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i \right\rangle_K \\ &= \left\langle f, \sum_{i=1}^{\infty} \lambda_i \langle e_i, K_x \rangle_K e_i \right\rangle_K \\ &= \sum_{i=1}^{\infty} \lambda_i e_i(x) \langle f, e_i \rangle_K \\ &= 0 \end{aligned}$$

where the last step holds because f is orthogonal to all e_i. This result shows that $\mathcal{H}_K$ is spanned by $\{e_i\}$. Therefore $\sqrt{\lambda_i} e_i$ is an orthonormal basis for $\mathcal{H}_K$.

Note: Another way of proving this result would be to consider the square root $I_K^{1/2}$ of the integral operator. $I_K^{1/2}$ is an isometry $L^2_\rho \to \mathcal{H}_K$, which is to say:

$$\langle f, g \rangle_\rho = \langle L_K^{1/2} f, L_K^{1/2} g \rangle_K, \qquad \forall f, g \in \mathcal{H}_K$$

And a unit-norm eigenbasis for $L_K^{1/2}$ is $\sqrt{\lambda_i} e_i$. □

Lemma A.3.2. *A function $f = \sum_{i=1}^{\infty} a_i e_i$ lies in the image of I_K if and only if*

$$\sum_{i=1}^{\infty} b_i^2 < \infty \tag{A.8}$$

where $b_i = a_i / \lambda_i$.

Proof. Suppose Equation A.8 holds. Let $g = \sum_{i=1}^{\infty} b_i e_i \in L^2_\rho$. Applying I_K

yields:

$$\begin{aligned} I_K(g) &= I_K(\sum_{i=1}^{\infty} b_i e_i) = \sum_{i=1}^{\infty} \lambda_i b_i e_i \\ &= \sum_{i=1}^{\infty} a_i e_i = f \end{aligned}$$

Then f lies in the span of I_K.

For the converse, suppose $f = I_K(g)$ for some $g \in L^2_\rho$. By the lemma above, we can write $g = \sum_{i=1}^{\infty} b_i e_i \in L^2_\rho$, so we have $\sum_{i=1}^{\infty} b_i < \infty$, which is Equation A.8. □

A.3.2 Appendix: The Closure of span k_x

Let $\mathcal{S}$, $\mathcal{H}_{K_\mathcal{M}}$ and $\mathcal{S}_\mathcal{M}$ be defined as in section 5.2 (Lemma 5.2.3).

Lemma A.3.3. $\mathcal{H}_{K_\mathcal{M}} = \mathcal{S}_\mathcal{M}$

Proof. Let $f_\mathcal{M}$ be an arbitrary function in $\mathcal{M}$. By the completeness of $\mathcal{H}_{K_\mathcal{M}}$, we can write $f_\mathcal{M} = \lim_{n\to\infty} f_\mathcal{M}^{(n)}$, where $f_\mathcal{M}^{(n)}$ lies in the span of the kernel functions: $f_\mathcal{M}^{(n)} = \sum_i a_i^{(n)} K_{\mathcal{M},x}$.

Let $f^{(n)}$ be the corresponding sequence in $\mathcal{H}_K$: $f^{(n)} = \sum_i a_i^{(n)} K_x$.

We see that $f^{(n)}$ is a Cauchy sequence because $\left\| f^{(n)} - f^{(k)} \right\|_K = \left\| f_\mathcal{M}^{(n)} - f_\mathcal{M}^{(n)} \right\|_{\mathcal{M}_K}$ and $f_{\mathcal{M}_K}^{(n)}$ converges. Then the limit $f = \lim_{n\to\infty} f^{(n)}$ exists and $f = f_\mathcal{M}$.

Therefore $\mathcal{H}_{K_\mathcal{M}} \subset \mathcal{S}_\mathcal{M}$, and the converse follows with the spaces swapped. □

Lemma A.3.4. *The complement of $\mathcal{S}$ is $\mathcal{S}^\perp = \{f \in \mathcal{H} : f(\mathcal{M}) = 0\}$.*

Proof. If $f \in \mathcal{S}^\perp$, then f vanishes on $\mathcal{M}$ because $f(x) = \langle k_x, f\rangle_K = 0$ for $x \in \mathcal{M}$.

Conversely, if $f(\mathcal{M}) = 0$, then $\langle k_x, f\rangle_K = f(x) = 0$ for every $x \in \mathcal{M}$ so f is orthogonal to the closure of $\text{span}\{k_x : x \in \mathcal{M}\}$. □

References

[1] Positive definiteness, reproducing kernel hilbert spaces and beyond. *Annals of Functional Analysis*, 2013.

[2] Anis Ben Abdessalem, Nikolaos Dervilis, David J Wagg, and Keith Worden. Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential monte carlo. *Frontiers in Built Environment*, 3:52, 2017.

[3] William N Anderson Jr and Thomas D Morley. Eigenvalues of the laplacian of a graph. *Linear and multilinear algebra*, 18(2):141–145, 1985.

[4] Andreas Argyriou, Mark Herbster, and Massimiliano Pontil. Combining graph laplacians for semi–supervised learning. In *Advances in Neural Information Processing Systems*, pages 67–74, 2006.

[5] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. ISSN 00029947. URL http://www.jstor.org/stable/1990404.

[6] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15509–15519, 2019.

[7] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52 – 72, 2007. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2006.07.001. URL http://www.sciencedirect.com/science/article/pii/S0885064X06000781.

[8] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[9] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[10] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. In *International Conference on Computational Learning Theory*, pages 486–500. Springer, 2005.

[11] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[12] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[13] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

[14] Ethan Bolker, Victor Guillemin, and Tara Holm. How is a graph like a manifold? *arXiv preprint math/0206103*, 2002.

[15] Vincent Borrelli, Said Jabrane, Francis Lazarus, and Boris Thibert. Flat tori in three-dimensional space and convex integration. *Proceedings of the National Academy of Sciences*, 2012. ISSN 0027-8424. doi: 10.1073/pnas.1118478109. URL https://www.pnas.org/content/early/2012/04/18/1118478109.

[16] Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p-laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 81–88, 2009.

[17] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[18] Yaiza Canzani. Analysis on manifolds via the laplacian.

[19] Rui Castro. Statistical learning theory lecture notes, 2018.

[20] O. Chapelle, B. Scholkopf, and A. Zien, Eds. *Semi-Supervised Learning*, volume 20. 3 2009. doi: 10.1109/TNN.2009.2015974.

[21] I. Chavel, B. Randol, and J. Dodziuk. *Eigenvalues in Riemannian Geometry*. ISSN. Elsevier Science, 1984. ISBN 9780080874340. URL https://books.google.com/books?id=0v1VfTWuKGgC.

[22] Fan RK Chung. Lectures on spectral graph theory. *CBMS Lectures, Fresno*, 6: 17–21, 1996.

[23] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Large scale transductive svms. *Journal of Machine Learning Research*, 7(Aug):1687–1712, 2006.

[24] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[25] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[27] Francesco Dinuzzo and Bernhard Schölkopf. The representer theorem for hilbert spaces: a necessary and sufficient condition. In *Advances in neural information processing systems*, pages 189–196, 2012.

[28] Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

[29] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 5 2019. ISSN 1558-0792. doi: 10.1109/msp.2018.2887284. URL http://dx.doi.org/10.1109/MSP.2018.2887284.

[30] 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.

[31] JC Ferreira and Valdir Antônio Menegatto. Positive definiteness, reproducing kernel hilbert spaces and beyond. *Annals of Functional Analysis*, 4(1), 2013.

[32] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[33] Jean Gallier. Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. 2016.

[34] A Gammerman, V Vovk, and V Vapnik. Learning by transduction. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 148–155, 1998.

[35] Bryan R Gibson, Timothy T Rogers, and Xiaojin Zhu. Human semi-supervised learning. *Topics in cognitive science*, 5(1):132–172, 2013.

[36] C. Godsil and G.F. Royle. *Algebraic Graph Theory*. Graduate Texts in Mathematics. Springer New York, 2013. ISBN 9781461301639. URL https://books.google.com/books?id=GeSPBAAAQBAJ.

[37] Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019.

[38] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[39] Carolyn Gordon, David L Webb, and Scott Wolpert. One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society*, 27(1):134–138, 1992.

[40] Willem H Haemers. Interlacing eigenvalues and graphs. *Linear Algebra and its applications*, 226(228):593–616, 1995.

[41] Sebastian Haeseler, Matthias Keller, Daniel Lenz, and Radoslaw Wojciechowski. Laplacians on infinite graphs: Dirichlet and neumann boundary conditions, 2011.

[42] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[43] Chaoqun Hong, Jun Yu, Jane You, Xuhui Chen, and Dapeng Tao. Multi-view ensemble manifold regularization for 3d object recognition. *Information sciences*, 320:395–405, 2015.

[44] Kwang In Kim, James Tompkin, Hanspeter Pfister, and Christian Theobalt. Local high-order regularization on data manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5473–5481, 2015.

[45] Stefan Ivanov and Dimiter Vassilev. The lichnerowicz and obata first eigenvalue theorems and the obata uniqueness result in the yamabe problem on cr and quaternionic contact manifolds. *Nonlinear Analysis*, 126:262–323, 2015.

[46] Sune K Jakobsen. Mutual information matrices are not always positive semidefinite. *IEEE Transactions on information theory*, 60(5):2694–2696, 2014.

[47] Mohammad Javaheri. Dirichlet problem on locally finite graphs. *Discrete Applied Mathematics*, 155(18):2496 – 2506, 2007. ISSN 0166-218X. doi: https://doi.org/10.1016/j.dam.2007.06.018. URL http://www.sciencedirect.com/science/article/pii/S0166218X07002296.

[48] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.

[49] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML 1999)*.

[50] Thorsten Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 290–297, 2003.

[51] Mark Kac. Can one hear the shape of a drum? *The american mathematical monthly*, 73(4P2):1–23, 1966.

[52] Chiu-Yen Kao, Rongjie Lai, and Braxton Osting. Maximization of laplace-beltrami eigenvalues on closed riemannian surfaces. *ESAIM Control Optimisation and Calculus of Variations*, 23(2):685–720, 2017. doi: 10.1051/cocv/2016008. URL https://app.dimensions.ai/details/publication/pub.1056952120.

[53] Kwang I Kim, Florian Steinke, and Matthias Hein. Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In *Advances in Neural Information Processing Systems*, pages 979–987, 2009.

[54] W. Kim and M. M. Crawford. Adaptive classification for hyperspectral image data using manifold regularization kernel machines. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4110–4121, 11 2010. ISSN 1558-0644. doi: 10.1109/TGRS.2010.2076287.

[55] R. Kraut. *The Cambridge Companion to Plato*. Cambridge Companions to Philosophy. Cambridge University Press, 1992. ISBN 9780521436106. URL https://books.google.com/books?id=QmmBpP41slwC.

[56] Kwok-Kun Kwong. Faber-krahn inequality, 9 2017. URL https://cuhkmath.wordpress.com/2017/09/09/faber-krahn-inequality/.

[57] Bracha Laufer-Goldshtein, Ronen Talmon, and Sharon Gannot. Semi-supervised sound source localization based on manifold regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(8):1393–1407, 2016.

[58] Jun Ling and Zhiqin Lu. Bounds of eigenvalues on riemannian manifolds. *ALM*, 10: 241–264.

[59] Weifeng Liu, Zheng-Jun Zha, Yanjiang Wang, Ke Lu, and Dacheng Tao. p-laplacian regularized sparse coding for human activity recognition. *IEEE Transactions on Industrial Electronics*, 63(8):5120–5129, 2016.

[60] M. Loeve. *Probability Theory I*. Graduate Texts in Mathematics. Springer New York, 1977. ISBN 9781468494648. URL https://books.google.com/books?id=L6vhBwAAQBAJ.

[61] Xueqi Ma and Weifeng Liu. Recent advances of manifold regularization. In *Manifolds II-Theory and Applications*. IntechOpen, 2018.

[62] Jonathan H Manton, Pierre-Olivier Amblard, et al. A primer on reproducing kernel hilbert spaces. *Foundations and Trends® in Signal Processing*, 8(1–2):1–126, 2015.

[63] Tatiana Mantuano. Discretization of riemannian manifolds applied to the hodge laplacian. *American journal of mathematics*, 130(6):1477–1508, 2008.

[64] Russell Merris. A note on laplacian graph eigenvalues. *Linear algebra and its applications*, 285(1-3):33–35, 1998.

[65] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993, 2018.

[66] Kevin P Murphy. *Machine learning: a probabilistic perspective*. 2012.

[67] Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *J. Mach. Learn. Res.*, 14(1):1229–1250, 5 2013. ISSN 1532-4435.

[68] Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016. doi: 10.1017/CBO9781316219232.

[69] Guilherme Porto and Luiz Emílio Allem. Eigenvalue interlacing in graphs. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 5(1), 2017.

[70] Delip Rao, David Yarowsky, and Chris Callison-Burch. Affinity measures based on the graph laplacian. In *Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing*, pages 41–48. Association for Computational Linguistics, 2008.

[71] Satish Rao and Benjamin Weitz. Lectures on combinatorial algorithms and data structures.

[72] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[73] Frigyes Riesz and Bela Sz Nagy. Functional analysis, ungar, new york, 1955. *RieszFunctional analysis1955*, 1990.

[74] Lorenzo Rosasco and Tomaso Poggio. *Machine Learning: A Regularization Approach*. 2017.

[75] Kevin Schlegel. When is there a representer theorem? *Journal of Global Optimization*, 74(2):401–415, 2019. ISSN 1573-2916. doi: 10.1007/s10898-019-00767-0. URL https://doi.org/10.1007/s10898-019-00767-0.

[76] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[77] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[78] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[79] Daniel Spielman. Spectral graph theory and its applications. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 29–38. IEEE, 2007.

[80] Daniel Spielman. Spectral graph theory. In *Combinatorial scientific computing*, number 18. Citeseer, 2012.

[81] He Sun. Lectures on algorithmic spectral graph theory. 2017.

[82] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[83] Luca Trevisan. The cheeger inequality in manifolds. 2013.

[84] W. T. Tutte. How to draw a graph. 1963.

[85] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*, 2019.

[86] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17 (4):395–416, 2007.

[87] Etienne Vouga. Laplace-beltrami: The swiss army knife of geometry processing. IGS Summer School 2016, 2016.

[88] Libo Weng, Fadi Dornaika, and Zhong Jin. Graph construction based on data self-representativeness and laplacian smoothness. *Neurocomputing*, 207:476–487, 2016.

[89] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[90] Xiaofei He, Shuicheng Yan, Yuxiao Hu, P. Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 3 2005. ISSN 1939-3539. doi: 10.1109/TPAMI.2005.55.

[91] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[92] Jian-Wu Xu, Puskal P Pokharel, Kyu-Hwa Jeong, and Jose C Principe. An explicit construction of a reproducing gaussian kernel hilbert space. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[93] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians, 2014.

[94] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *The IEEE International Conference on Computer Vision (ICCV)*, 10 2019.

[95] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, language technologies institute, 2005.

[96] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.