Error Analysis of Laplacian Eigenmaps for Semi-supervised Learning

Xueyuan Zhou
Department of Computer Science
University of Chicago

Nathan Srebro
Toyota Technological Institute
Chicago

Abstract

We study the error and sample complexity of semi-supervised learning by Laplacian Eignmaps at the limit of infinite unlabeled data. We provide a bound on the error, and show that it is controlled by the graph Laplacian regularizer. Our analysis also gives guidance to the choice of the number of eigenvectors k to use: when the data lies on a d-dimensional domain, the optimal choice of k is of order $(n/\log(n))^{\frac{d}{d+2}}$, yielding an asymptotic error rate of $(n/\log(n))^{-\frac{2}{2+d}}$.

1 Introduction

Graph Laplacian plays a central role in several popular methods for semi-supervised learning (SSL). One popular approach is graph Laplacian regularization (Zhu et al., 2003), where a predictor is regularized using a graph Laplacian based penalty term intuitively measuring the variability of the predictor with respect to the empirical distribution (variants of this approach include Zhou et al., 2004; Belkin et al., 2004). At the limit of infinite unlabeled data, with appropriate scaling and normalization (see details in Section 3), this penalty term converges to an appealing measure of variability with respect to a density:

$$J_p(f) = \int_{\Omega} \|\nabla f(x)\|^2 p(x)dx \tag{1}$$

where $f(x)$ is a real-valued function and $p(x)$ is the underlying probability density from which data is drawn. The measure $J_p(f)$ intuitively captures the complexity of function f in a density dependent way. One can then hope that if there exists a low error predictor f^* with low complexity $J_p(f^*)$, then regularizing $J_p(f)$ can allow us to learn $f^*(\cdot)$ with a sample complexity that depends on $J_p(f^*)$.

Unfortunately, it was recently shown by Nadler et al. (2009) that $J_p(f^*)$ by itself is *not* a good regularizer. In particular, in domains having dimensions greater then one, $J_p(f)$ is minimized by a function perfectly fitting the data and almost zero everywhere else, thus allowing no generalization. Accordingly, Laplacian based regularization alone is not well-behaved in the limit of infinite unlabeled data and instead yields nearly constant predictions.

An alternative approach to semi-supervised learning using graph Laplacians is to restrict training only to those functions spanned by the leading eigenvectors of the graph Laplacian (Belkin and Niyogi, 2004). This approach, known as "Laplacian Eigenmaps SSL", is well-behaved at the limit of infinite unlabeled data. In this paper we show a tight connection between Laplacian Eigenmaps SSL, using an appropriately normalized Laplacian, and the Laplacian penalty $J_p(f)$. In particular, we show that if there exists a low error predictor f^* with low $J_p(f^*)$, then although it cannot be learned by directly minimizing the graph Laplacian penalty, it *can* be learned using Laplacian Eigenmaps, with a sample complexity determined by $J_p(f^*)$. This means that even though direct regularization with the graph Laplacian penalty is ill-posed, it is still relevant as a complexity measure, and we can still guarantee learning of functions for which the graph Laplacian regularizer is finite by using the Laplacian Eigenmaps.

Our main technical result (Theorem 2 in Section 4.1), which is used in order to establish the above relationship, is a bound on the approximation error in the Laplacian Eigenmaps space in terms of $J_p(f)$. In particular, we show that, in the limit of infinite unlabeled data, any function f for which $J_p(f) < \infty$ can be approximated by a function f_k spanned by the k leading eigenfunctions of the Laplacian, with error:

$$\mathbf{E}_{x\sim p}\left[(f(x) - f_k(x))^2\right] \leq \frac{J_p(f)}{\lambda_{k+1}} \tag{2}$$

where λ_k is the k^{th} eigenvalue of the limit of a graph

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

Laplacian, i.e., the weighted Laplacian (Grigor'yan, 2006). Combining (2) with the existing estimation error analysis in a finite dimensional space, we show the desired learning guarantees in Section (5).

Our results apply to Laplacian Eigenmaps SSL with a non-standard Laplacian normalization, which is different from the graph Laplacian regularization suggested by Belkin and Niyogi (2004). In Section (3) we discuss why this particular normalization, but not other alternatives, might be preferable.

Our analysis also provides insights into the choice of the number of eigenvectors k of a graph Laplacian to use, i.e., the dimensionality of the Laplacian Eigenmaps space. When the intrinsic dimensionality of the support of $p(x)$ is d, the optimal choice of k scales as $(n/\log(n))^{d/(d+2)}$, where n is the number of labeled points. Note the sub-linear dependence on n (depending on the dimensionality) as opposed to the linear choice $k = n/5$ originally suggested. With this optimal choice of k, and with enough unlabeled points, Laplacian Eignmaps SSL achieves an integrated mean square error rate of $(n/\log(n))^{-2/(d+2)}$, which is up to a logarithmic factor asymptotically optimal for nonparametric regressions *in the d-dimensional intrinsic space*, see e.g., (Bickel and Li, 2007). The results can be further generalized to m times differentiable functions, where the optimal k scales as $(n/\log(n))^{d/(d+2m)}$, and the corresponding optimal error rate is $(n/\log(n))^{-2m/(2m+d)}$.

2 Setup

Consider an unknown source distribution over input-label pairs (x, y) where $x \in \mathbb{R}^N$ and $y \in \mathbb{R}$, with $-1 \leq y \leq 1$. The goal of SSL is to learn a good predictor $f(x)$ for y given a few labeled examples and many unlabeled examples. We consider SSL in a transductive setting, where we are given a labeled sample $(x_1, y_1), \ldots, (x_n, y_n)$ of n labeled points, and an unlabeled sample $x_{n+1}, \ldots, x_u$ of $u - n$ unlabeled points. The u pairs (x_i, y_i) are drawn i.i.d. from the source distribution. We denote the sequence of labeled points as X_L, the sequence of unlabeled points as X_U and both together as X. We denote the label sequences as $Y_L \in \mathbb{R}^n$, $Y_U \in \mathbb{R}^{u-n}$ and $Y \in \mathbb{R}^u$ accordingly. The task of transductive SSL is to estimate Y_U based on the given information from X_L, Y_L and X_U. Note that even though for convenience we write $f(\cdot)$ as a function, we do not actually output a function over $\mathbb{R}^N$, but rather just predictions $f(x_i)$ for points in the sample. It is thus more correct to think of $f(\cdot)$ as a u-dimensional vector $f(X) = (f(x_1), \ldots, f(x_u)) \in \mathbb{R}^u$. See (Bengio et al., 2004) for a discussion on out-of-sample extensions to Laplacian-based SSL.

We will use $\langle f, g \rangle_u = \frac{1}{u}\sum_{i=1}^u f(x_i)g(x_i)$ to refer to the empirical inner product over the data, with corresponding norm $\|f\|_u^2 = \langle f, f \rangle_u$. The continuous counterparts are $\langle f, g \rangle_{L^2(p)} = \int_\Omega f(x)g(x)p(x)dx$, and $\|f\|^2_{L^2(p)} = \int_\Omega |f(x)|^2 p(x)dx$. Let the gradient of f be ∇f, and

$$\|\nabla f(x)\|^2_{L^2(p)} = \int_\Omega \nabla f(x) \cdot \nabla f(x) p(x) dx$$

We assume that x follows a continuous distribution with smooth density $p(x)$ on a compact d-dimensional support $\Omega \subset \mathbb{R}^N$ with smooth boundary $\partial\Omega$. For simplicity assume that $p(x)$ is bounded from above and away from zero: $0 < a \leq p(x) \leq b < +\infty$.

2.1 Graph Laplacian

Consider a weighted graph G with u vertices corresponding to the labeled and unlabeled points $x_1, \ldots, x_u$, and edge weights w_{ij} measuring the similarity between x_i and x_j defined as:

$$w_{ij} = K(\frac{\|x_i - x_j\|^2}{4t}) = e^{-\frac{\|x_i - x_j\|^2}{4t}} \tag{3}$$

where t is a bandwidth parameter to be specified. Let $W \in \mathbb{R}^{u \times u}$ denote the weight matrix and let D be its diagonal degree matrix with $D_{ii} = \sum_j w_{ij}$.

We consider a two-step normalized graph Laplacian defined as follows: first, let $\tilde{W} = D^{-1/2}WD^{-1/2}$ be a normalization of the weight matrix, and denote by $\tilde{D}$ its diagonal degree matrix $\tilde{D}_{ii} = \sum_j \tilde{w}_{ij}$. The two-step random walk normalized Laplacian we use in this paper is then defined as:

$$\tilde{L}_r = I - \tilde{D}^{-1}\tilde{W} \tag{4}$$

This normalized graph Laplacian is from a one-parameter family of normalized graph Laplacians studied by (Hein et al., 2005; Coifman and Lafon, 2006). The normalization step allows us to control the weight in the limit of graph Laplacians as will be shown in section (3). See (Hein, 2005, Chapter 2) and the reference therein for further discussions of this one parameter family of normalized graph Laplacians which includes (4).

For a given weight matrix $\tilde{W}$, there are two other versions of graph Laplacian that are closely connected to $\tilde{L}_r$ and will be used later. One is the unnormalized graph Laplacian defined as

$$\tilde{L}_u = \tilde{D} - \tilde{W} \tag{5}$$

and the other is symmetric normalized graph Laplacian

$$\tilde{L}_s = \tilde{D}^{-1/2}\tilde{L}_u\tilde{D}^{-1/2} = I - \tilde{D}^{-1/2}\tilde{W}\tilde{D}^{-1/2} \tag{6}$$

The right eigenvectors of $\tilde{L}_r$ are the same as those in the generalized eigenfunctions problem as the following

$$\tilde{L}_u \phi_i = \lambda_i \tilde{D} \phi_i \tag{7}$$

It is easy to see that $\tilde{L}_r = \tilde{D}^{-1}\tilde{L}_u$. The eigenvectors of $\tilde{L}_s$ and the right eigenvectors of $\tilde{L}_r$ have a one to one mapping. If $v_{r,i}$ is a right eigenvector of $\tilde{L}_r$ with eigenvalue λ_i, then $v_{s,i} = \tilde{D}^{1/2} v_{r,i}$ is an eigenvector of $\tilde{L}_s$ with the same eigenvalue λ_i (von Luxburg, 2007).

2.2 Laplacian Eigenmaps SSL

The variant of Laplacian Eigenmaps SSL (Belkin and Niyogi, 2004, but with a modified normalization) we study proceeds as follows: Given the labeled and unlabeled data, and a parameter k, we first find the leading k right eigenvectors $v_{r,1}, \ldots, v_{r,k}$ of the two-step normalized Laplacian $\tilde{L}_r$, with the smallest eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$. Here and throughout we take the eigenvectors to be normalized s.t. $\|v_{r,j}\|_u = 1$. We then perform an ordinary (unregularized) least squares regression in the k-dimensional space obtained by the mapping $x_i \mapsto (v_{r,1}(x_i), v_{r,2}(x_i), \cdots, v_{r,k}(x_i))$, where, continuing to represent vectors as functions, $v_{r,j}(x_i)$ is the coordinate of $v_{r,j}$ corresponding to data point x_i. More explicitly, we find the least squares predictor

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} \beta_j v_{r,j}(x_i))^2$$

and predict $\hat{f}(x_i) = \sum_{j=1}^{k} \hat{\beta}_j v_{r,j}(x_i)$.

3 Limit of Infinite Unlabeled Data

In this section we consider the limit of the two-step normalized Laplacian $\tilde{L}_r$ (defined in equation 4) and of the graph Laplacian regularizer $f^T \tilde{L}_u f$ as the number of unlabeled points goes to infinity.

3.1 Graph Laplacian with Infinite Data

As the number of data points goes to infinite, the empirical Laplacian converges to a well defined weighted Laplacian (Hein et al., 2005; Coifman and Lafon, 2006; Belkin and Niyogi, 2008). When $t \to 0$ as $u \to \infty$ at an appropriate rate (Hein et al., 2005, Theorem 3), for any smooth function $f(x)$ and any $x \in \Omega/\partial\Omega$, up to a constant scaling[1]:

$$\frac{\tilde{L}_r f(x)}{t^{d/2+1}} \overset{\text{a.s.}}{\to} \tilde{\Delta}_r f(x) = -\Delta f(x) - \frac{1}{p(x)} \langle \nabla p(x), \nabla f(x) \rangle$$

where Δ is the regular Laplace operator defined as the following in $\mathbb{R}^N$:

$$\Delta = \sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2}.$$

Let ϕ_i be the i^{th} right eigenfunction of $\tilde{\Delta}_r$, and λ_i be the associated eigenvalue. Since the limit operator $\tilde{\Delta}_r$ includes a density dependent drifting term, the measure under which ϕ_i are orthogonal needs to be clarified. In order to do this, it is useful for us to consider its symmetric counterpart $\tilde{L}_s$. Since $\tilde{L}_s$ is symmetric, it has u real valued eigenvalues and the corresponding eigenvectors form an orthogonal basis of $\mathbb{R}^u$. Since all the eigenvector v_{s_i} are orthogonal, then by the one to one mapping $v_{s,i} = \tilde{D}^{1/2} v_{r,i}$ ($v_{r,i}$ is right eigenvector of $\tilde{L}_r$ and $v_{s,i}$ is the eigenvector of $\tilde{L}_s$), we have the following for $i \neq j$:

$$\langle v_{r,i}, v_{r,j} \rangle_u = \langle \tilde{D}^{-1/2} v_{s,i}, \tilde{D}^{-1/2} v_{s,j} \rangle_u = v_{s,i}^T \tilde{D}^{-1} v_{s,j}$$

That is, the eigenvectors of $\tilde{L}_r$ are only orthogonal w.r.t. the $\tilde{D}$-inner product. Therefore, by finding the limit of the degree function $\tilde{D}(i,i)$, we can obtain the weighting under which ϕ_i are orthogonal. Let

$$\tilde{d}_{t,u}(X_i) = \tilde{D}(i,i) = \sum_{j=1}^{u} \tilde{w}_{ij}$$

By (Hein, 2005, Proposition 2.33), up to a constant scaling, for all $x \in \Omega/\partial\Omega$:

$$\frac{1}{ut^{d/2}} \tilde{d}_{t,u}(x) \xrightarrow{\text{a.s.}} \tilde{d}(x) = 1 \tag{8}$$

From the orthogonality of eigenvectors of $\tilde{L}_s$, we can see that if ψ_i are the eigenfunctions of the limit operator of $\tilde{L}_s$, then $\int_\Omega \psi_i(x)\psi_j(x)p(x)dx = 0$. This means for eigenfunctions of the limit of $\tilde{L}_r$, $\int_\Omega \phi_i(x)\phi_j(x)\tilde{d}(x)p(x)dx = 0$. Since $\tilde{d}(x) = 1$, we can obtain the following orthogonality lemma for the right eigenfunctions of $\tilde{\Delta}_r$:

[1] Here and elsewhere, these convergences of a discrete vector, on the left hand side, to a continuous function, on the right hand side, should be interpreted as follows: fix $x \in \Omega/\partial\Omega$ and consider a sample of u points consisting of x and $u-1$ other points chosen i.i.d. (equivalently, condition on random samples which include x). The left hand side is then the random variable corresponding to the appropriate coordinate of $\tilde{L}_r f$, and we state its convergence, almost surely, to the value of the function $\tilde{\Delta}_r f$ at x.

Lemma 1. *Let $\phi_i(x)$ be the i^{th} right eigenfunction of $\tilde{\Delta}_r$, then for $i \neq j$*

$$\langle \phi_i, \phi_j \rangle_{L^2(p)} = \int_\Omega \phi_i(x)\phi_j(x)p(x)dx = 0$$

We will further normalize ϕ_i such that $\|\phi_i(x)\|_{L^2(p)} = \langle \phi_i, \phi_i \rangle_{L^2(p)} = 1$.

Note that the result (8) is not necessarily true for x on the boundary $\partial\Omega$, where the degree function $\tilde{d}(x)$ converges to a different constant. However, the limit $\tilde{d}(x)$ will be finite (Coifman and Lafon, 2006, Lemma 9). Since the boundary has zero measure, this does not affect the integral in Lemma 1.

Boundary Condition An important point that will be used later is that when the domain Ω has a non-empty smooth boundary, the right eigenfunctions ϕ_i of $\tilde{\Delta}_r$ automatically satisfy the Neumann boundary condition (Nadler et al., 2006):

$$\frac{\partial \phi_i(x)}{\partial \mathbf{n}} = 0 \tag{9}$$

where $\mathbf{n}$ is the normal direction to the boundary $\partial\Omega$ at x.

3.2 Limit of Graph Laplacian Regularizer

We also consider the limit of the graph Laplacian regularizer $f^T \tilde{L}_u f$ as the number of data points u goes to infinity. When $t \to 0$ at an appropriate rate as $u \to \infty$, then the (scaled) discrete regularizer converges to the appealing limit (Hein, 2005, Theorem 2.43):

$$\frac{1}{u^2 t^{d/2+1}} f^T \tilde{L}_u f \xrightarrow{\text{a.s.}} J_p(f) = \int_\Omega |\nabla f(x)|^2 p(x)dx \tag{10}$$

where in $\mathbb{R}^N$

$$|\nabla f(x)|^2 = \langle \nabla f(x), \nabla f(x) \rangle = \sum_{i=1}^N (\frac{\partial f(x)}{\partial x_i})^2$$

Note that in the limit the norm of the gradient is weighted by the density $p(x)$, rather then by the *squared* density $p^2(x)$, as is the case for unnormalized graph Laplacian $L_u = D - W$ shown by (Bousquet et al., 2004). This is because we choose the two-step normalized graph Laplacian.

3.3 Laplacian Eigenmap SSL with Infinite Unlabeled Data

As our interest here is in the behavior in the limit of infinite unlabeled data, we analyze here the Laplacian Eigenmaps SSL method based on the limit weighted Laplacian discussed above. That is, instead of relying on an unlabeled sample we consider the case where we have the marginal $p(x)$ and use the weighted Laplacian $\tilde{\Delta}_r$. Specifically, given the marginal $p(x)$, n labeled points $(x_1, y_1), \ldots, (x_n, y_n)$ and parameter k, we consider using the leading k right eigenfunctions $\phi_1, \ldots, \phi_k$ with the smallest eigenvalues $\lambda_1 \leq \cdots \leq \lambda_k$, and solving the following least squares problem:

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^n (y_i - \sum_{j=1}^k \beta_j \phi_j(x_i))^2.$$

Predictions are then given by

$$\hat{f}(x) = \sum_{j=1}^k \hat{\beta}_j \phi_j(x) \tag{11}$$

4 Error Analysis

In this section, we decompose the overall integrated mean squares error (IMSE) of Laplacian Eigenmaps SSL into two types of error, given infinite unlabeled data. Define Laplacian Eigenmaps space

$$S_k = \{f : f = \sum_{i=1}^k \alpha_i \phi_i, \alpha_i \in \mathbb{R}, |\alpha_i| < \infty\}$$

Firstly, we use a function $f_k \in S_k$ from the subspace spanned by the first k eigenfunctions to approximate the regression function $f \in C^1$, which potentially lives in an infinite dimensional space. This generates the finite dimension function approximation error between C^1 and S_k, i.e., $\inf_{f_k \in S_k} \|f - f_k\|^2_{L^2(p)}$.

Secondly, we also need to use least squares to estimate another function $\hat{f}_k$ to approximate f_k given n labeled points, which contributes another least squares error $\|\hat{f}_k - f_k\|^2_{L^2(p)}$, where $\hat{f}_k$ depends on n.

In terms of the IMSE we have:

$$\begin{aligned} & \mathbb{E}_{X_L,Y_L}[\|f - \hat{f}_k\|^2_{L^2(p)}] \\ = \ & \mathbb{E}_{X_L,Y_L}[\|f - f_k + f_k - \hat{f}_k\|^2_{L^2(p)}] \\ \leq \ & \mathbb{E}_{X_L,Y_L}[\|f - f_k\|^2_{L^2(p)}] + \mathbb{E}_{X_L,Y_L}[\|f_k - \hat{f}_k\|^2_{L^2(p)}] \end{aligned}$$

Function f and f_k are independent of sample data X_L and Y_L, while $\hat{f}_k$ depends on X_L and Y_L. Next we study the two types of error individually.

4.1 Approximation Error

For an arbitrary bounded smooth density, the approximation error between the spaces C^1 and S_k is given in the following theorem.

Theorem 2. *For $f \in C^1(\Omega)$ and a smooth density $p(x)$ such that $0 < a \le p(x) \le b < \infty$, for $k = 1, 2, \cdots$,*

$$\inf_{f_k \in S_k} \|f - f_k\|^2_{L^2(p)} \le \frac{J_p(f)}{\lambda_{k+1}} \qquad (12)$$

where λ_{k+1} is the $(k+1)^{th}$ eigenvalue of the weighted Laplacian $\tilde{\Delta}_r$.

Proof. For $f \in C^1(\Omega)$ on a compact domain Ω, $\|f\|^2_{L^2(p)} < \infty$ and $J_p(f) < \infty$. Since ϕ_i form an orthogonal basis w.r.t. $p(x)$, we can expand f as $f = \sum_{i=1}^{\infty} \alpha_i \phi_i$, where $\alpha_i = \langle f, \phi_i \rangle_{L^2(p)} = \int_\Omega f(x)\phi_i(x)p(x)dx$. Let $f_k = \sum_{i=1}^{k} \alpha_i \phi_i$ and

$$r_k = f - f_k$$

By the orthogonality of ϕ_i,

$$\langle r_k, \phi_i \rangle_{L^2(p)} = 0, \text{ for } i = 1, \cdots, k$$

Then by the minimum potential principle for λ_{k+1}:

$$\lambda_{k+1} = \min_{\substack{\langle g, \phi_i \rangle_{L^2(p)} = 0 \\ \text{for } i=1,\cdots,k}} \frac{\|\nabla g\|^2_{L^2(p)}}{\|g\|^2_{L^2(p)}} \le \frac{\|\nabla r_k\|^2_{L^2(p)}}{\|r_k\|^2_{L^2(p)}} \qquad (13)$$

Now consider the integral

$$\|\nabla r_k\|^2_{L^2(p)} = \int_\Omega |\nabla(f - f_k)|^2 p(x)dx$$
$$= \int_\Omega (|\nabla f(x)|^2 + |\nabla f_k(x)|^2 - 2\langle \nabla f, \nabla f_k \rangle)p(x)dx$$

For the last term, by the Green's identity

$$\int_\Omega \langle \nabla f, \nabla f_k \rangle p(x)dx$$
$$= \int_\Omega f(x)\tilde{\Delta}_r f_k(x)p(x)dx + \oint_{\partial\Omega} p(x)f(x)\nabla_{\mathbf{n}} f_k(x)dx$$
$$= \int_\Omega f(x)\tilde{\Delta}_r f_k(x)p(x)dx$$

The boundary integral vanishes since $f_k(x)$ is a finite linear combination of $\phi_i(x)$ so it satisfies the Neumann boundary condition. Then

$$\int_\Omega f(x)\tilde{\Delta}_r f_k(x)p(x)dx$$
$$= \int_\Omega [\textstyle\sum_{i=1}^{\infty} \alpha_i \phi_i(x)][\sum_{i=1}^{k} \alpha_i \lambda_i \phi_i(x)]p(x)dx$$
$$= \textstyle\sum_{i=1}^{k} \alpha_i^2 \lambda_i$$
$$= \int_\Omega f_k(x)\tilde{\Delta}_r f_k(x)p(x)dx$$
$$= \int_\Omega |\nabla f_k(x)|^2 p(x)dx$$

therefore

$$\|\nabla r_k\|^2_{L^2(p)} = \int_\Omega |\nabla(f - f_k)|^2 p(x)dx$$
$$= \int_\Omega (|\nabla f(x)|^2 - |\nabla f_k(x)|^2)p(x)dx$$
$$\le \int_\Omega |\nabla f(x)|^2 p(x)dx = \|\nabla f\|^2_{L^2(p)}$$

Together with equation (13),

$$\lambda_{k+1} \le \frac{\|\nabla f\|^2_{L^2(p)}}{\|r_k\|^2_{L^2(p)}} \Rightarrow \|r_k\|^2_{L^2(p)} \le \frac{\|\nabla f\|^2_{L^2(p)}}{\lambda_{k+1}} \qquad \square$$

From the proof we can see that f_k is just f with the tailing high frequency components cut off, or the output of a low pass filter with input being f. This is due to the completeness and orthogonality of Laplacian eigenfunctions.

The message from this theorem is that $J_p(f)$ together with λ_{k+1} determines the IMSE by using $f_k \in S_k$ to approximate $f \in C^1$. Even though if we use $J_p(f)$ as the limit of $f^T \tilde{L}_u f$ (or other unnormalized or normalized Laplacian) directly in the least squares problem, the solutions degenerate (Nadler et al., 2009), it can still be used as a complexity measure if we restrict solutions to be from a finite dimension space S_k. This also shows that choosing finite dimensional space S_k to approximate an infinite dimensional space is another way of regularization.

4.2 Least Squares Estimate Error

We now turn to the least squares estimation error $\mathbb{E}_{X,Y_L}[\|f_k - \hat{f}_k\|^2_{L^2(p)}]$. This is now a standard least squares prediction problem with n observed labeled in the k-dimensional linear space S_k (see e.g., (Györfi et al., 2002, Chapter 11) and (Lee et al., 2002)), and we can apply standard distribution independent learning guarantees. For $k \ge 1$, by (Lee et al., 2002):

$$\mathbb{E}_{X_L,Y_L}[\|f_k - \hat{f}_k\|^2_{L^2(p)}] \le C_{\mathcal{S}} k \log(\tfrac{n}{k})/n$$
$$\le C_{\mathcal{S}} k \log(n)/n$$

where $C_{\mathcal{S}}$ is a constant independent of k and n.

4.3 Error for Laplacian Eigenmaps SSL

The total error for Laplacian Eigenmaps SSL is bounded by the summation of the two types of error, which is

$$\mathbb{E}_{X,Y_L}[\|f - \hat{f}_k\|^2_{L^2(p)}]$$
$$\le \mathbb{E}_{X,Y_L}[\|f - f_k\|^2_{L^2(p)}] + \mathbb{E}_{X,Y_L}[\|f_k - \hat{f}_k\|^2_{L^2(p)}]$$
$$= \tfrac{J_p(f)}{\lambda_{k+1}} + \tfrac{C_{\mathcal{S}} k \log(n)}{n}$$

This error bound is for any integer $n \ge 1$ and $k \ge 1$. Next by balancing the two terms, we show the optimal asymptotic k, and the corresponding optimal error rate.

5 Optimal Error Rate

We now turn to studying the asymptotics of the Laplacian Eigenmaps SSL as the number of labeled points increases. Note that we still consider the number of unlabeled points as infinite, and refer to the Infinite Unlabeled Data Laplacian SSL, but now also consider the scaling of the error and the optimal choice of k as the numbered of labeled points n increases.

The next theorem shows the optimal number of Laplacian eigenfunctions that should be used, and the corresponding optimal IMSE.

Theorem 3. *Given infinite unlabeled points, for regression function $f \in C^1$, and the least squares estimator (11) on a compact domain Ω with intrinsic dimension d, the optimal k and the corresponding optimal integrated mean squares error rate are*

$$\begin{aligned} k^* = &\ \left(\tfrac{2J_p(f)}{dC_{\mathcal{W}}C_{\mathcal{S}}}\tfrac{n}{\log(n)}\right)^{\frac{d}{d+2}} \sim O\left(\left(\tfrac{n}{\log(n)}\right)^{\frac{d}{d+2}}\right) \\ \mathit{IMSE}^* = &\ \left(\tfrac{n}{\log(n)}\right)^{-\frac{2}{d+2}}\left[\left(\tfrac{J_p(f)}{C_{\mathcal{W}}}\right)\left(\tfrac{2J_p(f)}{dC_{\mathcal{W}}C_{\mathcal{S}}}\right)^{-\frac{2}{d+2}} + \right. \\ &\ \left. C_{\mathcal{S}}\left(\tfrac{2J_p(f)}{dC_{\mathcal{W}}C_{\mathcal{S}}}\right)^{\frac{d}{d+2}}\right] \\ \sim &\ O\left(\left(\tfrac{n}{\log(n)}\right)^{-\frac{2}{2+d}}\right) \end{aligned} \tag{14}$$

where $C_{\mathcal{W}}$ is the Weyl constant defined by $C_{\mathcal{W}} = C_{\mathcal{W}}(d, Vol(\Omega)) = \frac{d}{d+1}\frac{4\pi^2}{(\omega_d\, Vol(\Omega))^{2/d}}$ and ω_d is the volume of a unit ball in $\mathbb{R}^d$. $C_{\mathcal{S}}$ is the constant coefficient from least squares estimate error bound of f.

Proof. From the following total error we can see that, in order to study the asymptotic behavior, we need to find the asymptotics for λ_k.

$$\frac{J_p(f)}{\lambda_{k+1}} + \frac{C_{\mathcal{S}}k\log(n)}{n}$$

By the Weyl's asymptotic formula for eigenvalues of Laplacian Δ on either $\mathbb{R}^d$ or a manifold with intrinsic dimension d (Safarov and Vassiliev, 1997), we have $\lambda_k \sim C_{\mathcal{W}}k^{\frac{2}{d}}$. For weighted Laplacian $\tilde{\Delta}_r$ with arbitrary smooth bounded density, we can not use the Weyl's formula directly. However, for a self-adjoint differential operator, the leading term of the asymptotics of its eigenvalues is determined by the highest order differential operator term of this operator (Safarov and Vassiliev, 1997). For a self-adjoint weighted Laplacian $\tilde{\Delta}_r = -\Delta - \frac{1}{p}\langle\nabla p, \nabla\rangle$ with a smooth and bounded density $p(x)$, the highest order differential operator is just the regular Laplacian, then the asymptotic for the eigenvalues of $\tilde{\Delta}_r$ is the same as

$$\lambda_k \sim C_{\mathcal{W}}k^{\frac{2}{d}}$$

Therefore, either on a compact domain of $\mathbb{R}^d$ or a d dimensional manifold, with arbitrary bounded smooth density, we have

$$\mathbb{E}_{X_L,Y_L}[\|f - \hat{f}_k\|^2_{L^2(p)}] \leq \frac{J_p(f)}{C_{\mathcal{W}}k^{\frac{2}{d}}} + \frac{C_{\mathcal{S}}k\log(n)}{n}$$

By balancing the two terms as a function of k, we can find the minimum by taking derivatives w.r.t. k to obtain the optimal k. Plugging the optimal k into the overall error leads to the optimal IMSE error rate. □

This theorem provides insights into the choice of the number of eigenvectors k of the graph Laplacian to use. When the intrinsic dimension of Ω is d, the optimal choice of k scales as $(n/\log(n))^{d/(d+2)}$ where n is the number of labeled points. Note the sub-linear dependence on n (also depending on the dimensionality) as opposed to the linear choice $k = n/5$ originally suggested.

Interestingly, with this optimal choice of k, and with infinite unlabeled points, Laplacian Eignmaps SSL achieves the same IMSE rate of $n^{-2/(d+2)}$ as the asymptotically optimal rate for nonparametric regressions *in the d-dimensional intrinsic space*, up to a logarithmic factor.

In this theorem, IMSE is considered. By replacing the IMSE with conditional mean square error (MSE) on X_L, the optimal k and the conditional MSE rate are

$$\begin{aligned} k^* \sim &\ O(n^{\frac{d}{d+2}}) \\ \text{MSE}^* \sim &\ O(n^{-\frac{2}{2+d}}) \end{aligned} \tag{15}$$

In this case, the optimal conditional MSE achieves the optimal rate as local polynomials *on the d-dimensional unknown manifolds* (Bickel and Li, 2007).

6 Analysis for C^m Functions

For simplicity, we assume the underlying probability density is uniform in this section. Based on above analysis we can see that the approximation error bound can be easily generalized to m times differentiable functions. For function $f \in C^m$, define

$$\begin{aligned} J^d_m(f) = &\ \textstyle\sum_{|\alpha|=m}\frac{m!}{\alpha_1!\cdots\alpha_d!}\int_\Omega\left(\frac{\partial^m f(x)}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right)^2 dx \\ = &\ \textstyle\sum_{|\alpha|=m}\binom{m}{\alpha}\|D^m f\|^2_{L^2} \end{aligned}$$

where

$$D^\alpha f = \frac{\partial^{|\alpha|}f}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\cdots\partial x_d^{\alpha_d}}$$

By the following classic relation, see e.g., (Chapter 6, Berlinet and Thomas-Agnan, 2003)

$$\Delta^m = \sum_{|\alpha|=m}\binom{m}{\alpha}D^{2\alpha}$$

with proper boundary conditions, we can obtain the following important relation

$$J_m^d(f) = \int_\Omega f(x)\Delta^m f(x)dx$$

The corresponding semi-norm in Fourier basis form is $\sum_{i=1}^\infty \alpha_i^2 \lambda_i^m$, where $\alpha_i = \langle f, \phi_i \rangle_{L^2(p)}$, see e.g., (Taylor, 1996). Since λ_i is increasing, then it is not difficult to see that the first error, finite dimension function approximation error becomes

$$\inf_{f_k \in S_k} \|f - f_k\|_{L^2(p)}^2 \leq \frac{J_m^d(f)}{\lambda_{k+1}^m}$$

Since Δ and Δ^m share the same eigenfunctions, and when λ_i is the eigenvalue of Δ, λ_i^m is the eigenvalue of Δ^m. Next theorem generalizes results of theorem (3) from C^1 functions to C^m functions

Theorem 4. *Given infinite unlabeled points, for regression function $f \in C^m$, and the least squares estimator (11)on a compact domain Ω with intrinsic dimension d, the optimal k and the corresponding optimal mean squares error rate are*

$$\begin{aligned} k^* &= (\tfrac{2mJ_m^d(f)}{dC_{\mathcal{W}}C_{\mathcal{S}}}\tfrac{n}{\log(n)})^{\frac{d}{d+2m}} \sim O((\tfrac{n}{\log(n)})^{\frac{d}{d+2m}}) \\ MSE^* &= (\tfrac{n}{\log(n)})^{-\frac{2m}{d+2m}}[\tfrac{J_m^d(f)}{C_{\mathcal{W}}})(\tfrac{2mJ_m^d(f)}{dC_{\mathcal{W}}C_{\mathcal{S}}})^{-\frac{2m}{d+2m}} + \\ &\qquad C_{\mathcal{S}}(\tfrac{2mJ_m^d(f)}{dC_{\mathcal{W}}C_{\mathcal{S}}})^{\frac{d}{d+2m}}] \\ &\sim O((\tfrac{n}{\log(n)})^{-\frac{2m}{2m+d}}) \end{aligned} \tag{16}$$

where $C_{\mathcal{W}}$ is the Weyl constant defined by $C_{\mathcal{W}} = C_{\mathcal{W}}(d, Vol(\Omega)) = \frac{d}{d+1}\frac{4\pi^2}{(\omega_d Vol(\Omega))^{2/d}}$ and ω_d is the volume of unit ball in $\mathbb{R}^d$. $C_{\mathcal{S}}$ is the constant coefficient from least squares estimate error bound of f.

This theorem generalizes previous results to even smoother functions. The IMSE rate, up to a logarithmic factor, achieves the optimal error rate of nonparametric regressions for C^m functions. Compared to the results of $m = 1$, if m is large, the optimal k will scale slower than $m = 1$ case. We can possibly use much less eigenfunctions to achieve the optimal error. Similar results also hold for the conditional MSE.

7 Discussion

Comparison to Local Polynomial Regression

Local polynomial regression is recently shown to be able to achieve the same conditional MSE convergence rate $n^{-\frac{2m}{2m+d}}$ for m times differentiable functions on a d-dimensional manifold, without estimating the manifold (Bickel and Li, 2007). This is not surprising considering "locally, the geodesic distance is roughly proportionate to the Euclidean distance". Compared to Laplacian Eigenmaps SSL, the most important advantage of local polynomial regression is that it does not need to estimate the manifold.

However, the cost of avoiding the estimation of manifolds is the optimal window bandwidth selection and local dimension estimation. For relatively large d, with extremely sparse data points, either of these problems can be easily solved. Moreover, for larger m, free parameters to be estimated for polynomials increase fast as d increases, which adds another difficulty for it to be a practical algorithm. Another difference is that local polynomial regression is a supervised learning algorithm, while Laplacian Eigenmaps SSL is a transductive SSL algorithm using data dependent basis.

Comparison to Regression Splines in Finite Dimension Subspaces

Regression using the first k eigenfunctions can be seen as approximating functions in finite dimensional subspaces, which can be compared to regression splines in finite dimension subspaces, see e.g., (Wahba, 1990, Chapter 7). When $d = 1$, our optimal $k \sim O(n^{\frac{1}{1+2m}})$ is the same as splines regression over 1-dimensional intervals, and achieves the same IMSE (both up to $\log(n)$), $O(n^{-\frac{2m}{2m+1}})$, as shown by Agarwal and Studden (1980).

Our analysis also shows that Laplacian Eigenmaps SSL achieves the same error rate using the same optimal dimensions compared to least squares estimates using tensor product spline spaces with equidistant knots (Chapter 15.3, Györfi et al., 2002).

Laplacian Eigenmaps Space

Given Laplacian Eigenmaps space S_k, if we use another learning algorithm in this k-dimensional space that minimizes the mean squares error, we can plug in the corresponding error rates to obtain the optimal dimension k and error rate, which can be potentially improved by a logarithmic factor, achieving the exact optimal nonparametric regression rate.

Normalized Graph Laplacian

In this paper, we use a two-step normalized graph Laplacian. In fact, there is a family of one parameter normalized graph Laplacian (Coifman and Lafon, 2006; Hein et al., 2005), which is defined as follows: first normalize weight matrix as $\tilde{W}_\alpha = D^{-\alpha}WD^{-\alpha}$ for $\alpha \in \mathbb{R}$, then the graph Laplacian is $\tilde{L}_\alpha = I - \tilde{D}_\alpha^{-1}\tilde{W}_\alpha$. Similarly, for each α, there are unnormalized, random walk normalized and symmetric normalized graph Laplacians. Then a similar analysis follows easily. The only difference is that the weight measure is not $p(x)$ anymore, instead, it will be $[p(x)]^{2-2\alpha}$.

Finite Unlabeled Points Analysis

In this paper, we considered the "Laplacian Eigenmaps SSL with Infinite Unlabeled Data" where we project the data onto the leading eigenfunctions of the limit operator of a normalized graph Laplacian. This leaves open an important issue of the convergence of the Laplacian Eigenmaps to this infinite data limit, and in particular the rate of this convergence. Although convergence of the *operator* is well understood, we are not aware of a rigorous analysis for the convergence rate of the eigenvectors of a graph Laplacian to the eigenfunctions of its limit operator. Such an analysis is crucial for applying the results obtained here to obtain specific guarantees with finite amounts of unlabeled data. Nevertheless, we consider this "infinite unlabeled data" analysis useful at understanding the Laplacian Eigenmaps SSL method, and point out that a convergence result on the eigenvectors, when obtained, could be combined with our analysis to obtain a full understanding of Laplacian Eigenmaps SSL error rate as a function of the number of labeled and unlabeled points.

Acknowledgments

We thank the anonymous reviewers for helping to improve the paper.

References

G. Agarwal and W. Studden. Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist*, 8(6):1307–1325, 1980.

M. Belkin and P. Niyogi. Towards a Theoretical Foundation for Laplacian-Based Manifold Methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

M. Belkin and P. Niyogi. Semi-supervised Learning on Riemannian Manifolds. *Machine Learning, Special Issue on Clustering*, 56:209–239, 2004.

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and Semi-supervised Learning on Large Graphs. In John Shawe-Taylor and Yoram Singer, editors, *COLT*, volume 3120 of *Lecture Notes in Computer Science*, pages 624–638. Springer, 2004.

Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers, 2003.

Peter J. Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond, IMS Lecture Notes-Monograph Series*, 54:177–186, 2007.

Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure Based Regularization. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

A. Grigor'yan. Heat kernels on weighted manifolds and applications. *Cont. Math.*, 398:93–191, 2006.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression.* Springer, 2002.

M. Hein. *Geometrical aspects of statistical learning theory.* PhD thesis, Wissenschaftlicher Mitarbeiter am Max-Planck-Institut für biologische Kybernetik in Tübingen in der Abteilung, 2005.

M. Hein, J. Audibert, and U. von Luxburg. From graphs to manifolds: Weak and strong pointwise consistency of graph Laplacians. In *Proc. 18th Conf. Learning Theory (COLT), Lecture Notes Comput. Sci.*, volume 3559. Springer-Verlag, Berlin, 2005.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *Information Theory, IEEE Transactions on*, 44(5):1974–1980, 2002. ISSN 0018-9448.

Boaz Nadler, Stephane Lafon, Ronald Coifman, and Ioannis Kevrekidis. Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 955–962, Cambridge, MA, 2006. MIT Press.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Statistical Analysis of Semi-Supervised Learning: The Limit of Infinite Unlabelled Data. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1330–1338, 2009.

Yu. Safarov and D. Vassiliev. *The Asymptotic Distribution of Eigenvalues of Partial Differential Operators*, volume 155. American Mathematical Society, 1997.

Michael E. Taylor. *Partial Differential Equations I: Basic Theory.* Springer, New York, 1996.

U. von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416, 2007.

G. Wahba. *Spline Models for Observational Data.* Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: SIAM, 1990.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with Local and Global Consistency. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.

X. Zhu, Ghahramani, Z., and J. Lafferty. Semi-Supervised Learning Using Gaussian Fields and Harmonic Function. In *The Twentieth International Conference on Machine Learning*, 2003.