Semi-supervised Learning by Higher Order Regularization

Xueyuan Zhou
Department of Computer Science
University of Chicago

Mikhail Belkin
Department of Computer Science and Engineering
The Ohio State University

Abstract

In semi-supervised learning, at the limit of infinite unlabeled points while fixing labeled ones, the solutions of several graph Laplacian regularization based algorithms were shown by Nadler et al. (2009) to degenerate to constant functions with "spikes" at labeled points in $\mathbb{R}^d$ for $d \geq 2$. These optimization problems all use the graph Laplacian regularizer as a common penalty term.

In this paper, we address this problem by using regularization based on an iterated Laplacian, which is equivalent to a higher order Sobolev semi-norm. Alternatively, it can be viewed as a generalization of the thin plate spline to an unknown submanifold in high dimensions. We also discuss relationships between Reproducing Kernel Hilbert Spaces and Green's functions. Experimental results support our analysis by showing consistently improved results using iterated Laplacians.

1 Introduction

Graph Laplacian regularization is one of the most popular semi-supervised learning (SSL) methods, see e.g., (Chapelle et al., 2006) and (Zhu, 2008). Several SSL methods solve optimization problems by penalizing regularizer $f^T Lf$ (or its variations), whose limit given infinite data points is an "energy" term

$$\int_\Omega \|\nabla f(x)\| p^\alpha(x) dx$$

where L is the graph Laplacian, $p(x)$ is the underlying probability density and α is a non-negative number. A typical form of the optimization problem (Zhu et al., 2003; Zhou et al., 2004; Belkin et al., 2004) is

$$\min_f \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu f^T Lf \qquad (1)$$

where $f^T Lf = \frac{1}{2}\sum_{x_i,x_j \in X} w_{ij}(f(x_i) - f(x_j))^2$, $X = X_L \cup X_U$, X_L is the sample set whose label y_i is known and X_U the unlabeled sample set whose labels are unknown, and w_{ij} is a similarity weight between sample x_i and x_j.

However, in the limit of infinite unlabeled data while holding the labeled points fixed, with a proper scaling of $\mu \to 0$, solutions to problem (1) degenerate to constant functions on unlabeled points, with "spikes" at labeled points (Nadler et al., 2009). This shows there is no generalization in the limit, and also results in unstable solutions in finite unlabeled data case.

In this paper, we show that regularization using higher order Sobolev semi-norm $\|f\|_m$ of order m resolves this problem, given $2m > d$ where d is the intrinsic dimension of the submanifold. The Sobolev embedding theorem, e.g. see (Adams, 1975, Chapter 5), guarantees that this method gives continuous solutions of a certain order in the limit of infinite unlabeled points while fixing labeled ones. We use iterated Laplacian semi-norm $\int_\Omega f(x)\Delta^m f(x)dx$ as this Sobolev semi-norm, which corresponds to the empirical iterated Laplacian regularizer $f^T L^m f$ given finite data, and also has the advantage of being coordinate free.

Iterated Laplacian regularizer $f^T L^m f$ or its variation $f^T g(L)f$ have been used by (Smola and Kondor, 2003; Belkin et al., 2004, 2006). Unlike previous works, we focus on the analysis from a Sobolev space point of view given infinite unlabeled points, and in particular, study the condition for regularizer $f^T L^m f$ to give continuous solutions. We note that some Laplacian-based methods, e.g., (Belkin et al., 2006) are guaranteed to provide continuous solutions due to a different formulation of the optimization problem.

A parallel relation between Green's functions and reproducing kernels based on the graph Laplacian also

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.

plays an important role in this problem, see e.g., (Wahba, 1990; Ramsay and Silverman, 1997). By comparing Green's functions and reproducing kernels, we not only can explain the degenerate solution in problem (1), but also can explore a close relation between the Sobolev space and the Hilbert reproducing kernel space (RKHS) in a more general setting as well as connections between graph Laplacian regularization and kernel methods.

In practice, SSL by iterated Laplacian regularization requires only a trivial modification of the code incorporating a power of the Laplacian matrix. The experimental results support our analysis by pointing to consistent and often significant improvements in classification error resulting from using iterated Laplacians. These improvements appear both on simulated data obtained from a mixture of Gaussian distributions and a number of standard datasets.

1.1 Problem Setup

Let $p(x)$ be a fixed unknown smooth probability density on a compact connected submanifold $\Omega \subset \mathbb{R}^N$ with boundary $\partial\Omega$, which can be empty. Let the intrinsic dimension of Ω be $d \leq N$, $0 < a \leq p(x) \leq b < +\infty$, $p(x)$ be infinitely differentiable for simplicity, and $f : \Omega \to \mathbb{R}$ be the unknown function to be estimated. By convenient abuse of notation, $f(x)$ means both the continuous function value at x and the x element of column vector f such that $f_x = f(x)$.

The task of SSL is to estimate $f(x)$ given l pairs of labeled points $\{(x_1, y_1) \cdots (x_l, y_l)\}$, and u unlabeled points $\{x_{l+1}, \cdots, x_{l+u}\}$, where $x_i \in \Omega$ is drawn from $p(x)$ i.i.d. y_i is the observed function value at x_i. We denote the labeled data set as $X_L = \{x_1, \cdots, x_l\}$, the unlabeled data set as $X_U = \{x_{l+1}, \cdots, x_{l+u}\}$, $X = X_L \cup X_U$, the corresponding label sets as Y_L and Y_U. Let $n = l + u$ be the total number of sample points. Ideally we want to estimate $f(x)$ on the whole submanifold Ω. In a transductive setting, instead, we estimate $f(x)$ on X_U (or Y_U), given X_U, X_L and Y_L. An interesting case is when Ω is a d-dimensional submanifold of Euclidean space $\mathbb{R}^N$ such that $d << N$.

In graph Laplacian SSL, all data points are mapped into vertices of an undirected graph $G(V, E)$. Given l labeled data points and u unlabeled points $(x_1, \cdots, x_{l+u})$, we build a weighted undirected graph $G(V, E)$ such that x_i is mapped to vertex V_i, and edge weight $w(e_{ij}) = w_{ij}$ is a similarity measure between x_i and x_j. Denote the connection weight matrix on graphs as W. A typical weight function is $w_{ij} = e^{-\|x_i - x_j\|^2/t}$. Let D be a diagonal matrix with $D_{ii} = \sum_j w_{ij}$, then $L = D - W$ is the unnormalized graph Laplacian. There are several versions of normalized graph Laplacian (von Luxburg, 2007; Coifman and Lafon, 2006), which we will compare in the experiment. A continuous Laplacian in $\mathbb{R}^d$ is defined as

$$\Delta = -\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$$

The limit of $f^T L f$ for a fixed function f as $n \to \infty$ and $t \to 0$ (Bousquet et al., 2004) at a specific rate is

$$\frac{1}{n^2 t^{d/2+1}} f^T L f \xrightarrow{p} \int_\Omega \|\nabla f\|^2 p^2(x) dx$$

where $\|\nabla f(x)\|^2 = \langle \nabla f(x), \nabla f(x) \rangle = \sum_{i=1}^{d} (\frac{\partial f(x)}{\partial x_i})^2$ in $\mathbb{R}^d$. Then by decreasing μ as $\mu / n^2 t^{d/2+1}$ which tends to zero, minimization problem (1) given infinite unlabeled points becomes

$$\min_f \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu \int_\Omega \|\nabla f(x)\|^2 p^2(x) dx \quad (2)$$

1.2 Problem in Higher Dimensions

When $d = 1$, problem (2) has a reasonable continuous solution. The solution space is an RKHS, the kernel of which is a density weighted Mexican hat function as shown by Nadler et al. (2009). When $d \geq 2$, the solution space is too large to give a continuous solution, and in fact the solution is a constant function on X_U, with "spikes" of value Y_L on X_L. This can also be seen as overfitting to the data Y_L since the solution function space is too rich.

Intuitively, the problem comes from the integral of gradient square in higher dimensions. Since the volume $dx = dx_1 dx_2 \cdots dx_d$ is small in higher dimensions, it allows $\|\nabla f(x)\|^2$ to be $+\infty$, while the whole integral $\int_\Omega \|\nabla f(x)\|^2 dx$ can still be small, or even zero as long as the infinitesimal quantity $dx_1 dx_2 \cdots dx_d$ has a higher order.

This problem can also be explained from the point of view of unbounded Green's functions as discussed later in the paper. This phenomenon is well known in splines, see e.g., (Wahba, 1990), but has drawn less attention in graph Laplacian regularization SSL community, which is possibly due to the fact that the finiteness of the data in practical applications can be viewed as an additional regularizer. Still, the experimental results show marked improvements when this issue is addressed as we show in Section 5.

2 SSL by Higher Order Regularization

In this section, we introduce one solution to this problem by higher order Sobolev semi-norm regularization, which can be implemented by an iterated Laplacian semi-norm. We first review several basic facts about Sobolev spaces, see e.g., (Adams, 1975).

2.1 Sobolev Space Review

Only real-valued functions are considered in this paper. Let $\mathbb{Z}^d_+$ denote the set of all ordered d-tuples of nonnegative integers. For $\alpha \in \mathbb{Z}^d_+$, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$, each component α_i being a nonnegative integer. We denote $|\alpha| = \sum_{i=1}^d \alpha_i$ and by $D^\alpha f$ the partial derivative

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

A Sobolev space of order m, denoted by $H^m(\Omega)$ (or $W^{m,2}(\Omega)$) is defined to be the space consisting of those functions in $L^2(\Omega)$ that, together with all their weak partial derivatives up to and including those of order m, belong to $L^2(\Omega)$

$$H^m(\Omega) = \{f : D^\alpha f \in L^2(\Omega)\ \forall \alpha \text{ s.t. } |\alpha| \leq m\}$$

See (Adams, 1975) for more general Sobolev Space $W^{m,p}(\Omega)$. $H^m(\Omega)$ is frequently used in applications to boundary value problems, see e.g., (Reddy, 1997). Define Sobolev semi-inner product $\langle u, v\rangle_m$ which consists of derivatives of order only m as

$$\sum_{|\alpha|=m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \int_\Omega (\frac{\partial^m u(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}})(\frac{\partial^m v(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}})dx$$

The induced Sobolev semi-norm, see e.g., (Berlinet and Thomas-Agnan, 2003, Chapter 6), is

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

Notice that $J^d_m(f)$ is used in thin plate splines on $\mathbb{R}^d$ as a regularizer, see e.g., (Wahba, 1990). In subspace orthogonal to its null, $J^d_m(f)$ is a norm.

Denote by C^k a function class whose derivatives up to order k are continuous. Next theorem describes an important relation between spaces $H^m(\Omega)$ and C^k, see e.g., (Adams, 1975). The theorem also provides a direct solution to our SSL problem in high dimensions.

Theorem 1. *(The Sobolev Embedding Theorem) Let Ω be a bounded domain in $\mathbb{R}^d$ with a Lipschitz boundary $\partial\Omega$. If $m - \frac{d}{2} > k$, then*

$$H^m(\Omega) \to C^k(\overline{\Omega})$$

where k is a nonnegative integer.

By abuse of notation, we can also write $H^m(\Omega) \subseteq C^k(\overline{\Omega})$ for an intuitive understanding. Since $H^m(\Omega)$ is really equivalent classes of functions up to sets measure of zero, the meaning of the embedding, or "$\subseteq$" has to be clarified. $H^m(\Omega) \to C^k(\overline{\Omega})$ means that each $u \in H^m(\Omega)$ can be modified on sets of zero measure to get $\tilde{u}$ in such a way that $\tilde{u} \in C^k(\overline{\Omega})$, and $\|\tilde{u}\|_{C^k(\overline{\Omega})} \leq c\|u\|_{H^m(\Omega)}$, with c being a constant.

Based on the Sobolev embedding theorem, the following theorem can be found in (Adams, 1975) and (Berlinet and Thomas-Agnan, 2003, Appendix), which connects space $H^m(\mathbb{R}^d)$ and an RKHS.

Theorem 2. *$H^m(\mathbb{R}^d)$ is an RKHS iff $2m > d$.*

2.2 Iterative Laplacian Semi-Norm

Define the iterated Laplacian semi-norm as

$$I^d_m(f) = \int_\Omega f(x) \Delta^m f(x) dx \tag{3}$$

and its empirical version as

$$I^d_{m,n}(f) = f^T L^m f$$

where n means L is built on total n data points. We show in next section that $I^d_{m,n}(f)$ converges to $I^d_m(f)$ in probability for a smooth function f, either on submanifolds without boundaries, or on submanifolds with boundaries given proper boundary conditions. Notice that since L is a real symmetric positive semi-definite matrix, $I^d_{m,n}(f)$ is a semi-norm without further conditions. The null space is spanned by the first eigenvector since $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. However, we need proper boundary conditions for $I^d_m(f)$ to be a semi-norm when the boundary set is not empty. Next lemma shows the conditions used in this paper for $I^d_m(f)$ to be a semi-norm

Lemma 3. *Given one of the following two conditions, $I^d_m(f) \geq 0$ and*

$$I^d_m(f) = \left\{ \begin{array}{l} \int_\Omega [\Delta^{\frac{m}{2}} f(x)]^2 dx, \ \textit{even } m \\ \int_\Omega \|\nabla[\Delta^{\frac{m-1}{2}} f(x)]\|^2 dx, \ \textit{odd } m \end{array} \right. \tag{4}$$

1. *Submanifold Ω has no boundary, i.e., $\partial\Omega = \emptyset$*
2. *$\nabla(\Delta^k f(x)) \cdot \mathbf{n} = 0$ for $x \in \partial\Omega$, $k = 0, \cdots, m-1$*

where $\mathbf{n}$ is the normal direction on the boundary and Δ^0 is the identity operator.

Equation (4) follows from iterative applications of the Green's identity. In fact, $I^d_m(f)$ is closely related to $J^d_m(f)$, which is used in thin plate splines, see e.g., (Berlinet and Thomas-Agnan, 2003, Chapter 6), (Wahba, 1990) and (Taylor, 1996).

An alternative useful way of defining Sobolev space $H^m(\Omega)$ is through Fourier basis, see e.g., (Taylor, 1996, Chapter 4 and Chapter 5 App. A), which is not only coordinate free, but also closely related to the iterated Laplacian semi-norm. Let the i^{th} eigenvalue and eigenfunction of the self-adjoint Δ be λ_i and $\phi_i(x)$. Since the eigenfunctions $\phi_i(x)$ form an $L^2(\Omega)$ basis, for a given $f \in L^2(\Omega)$, $f = \sum_{i=1}^{\infty} \hat{f}_i \phi_i, \hat{f}_i = \langle f, \phi_i \rangle_{L^2}$, where $\langle u, v\rangle_{L^2} = \int_\Omega u(x)v(x)dx$. For $s \geq 0$, define

$$D_s = \{f \in L^2(\Omega) : \sum_{i=1}^{\infty} |\hat{f}_i|^2 \lambda_i^s < \infty\} \tag{5}$$

then by (Taylor, 1996, Chapter 5 (A.18))

$$D_s \subset H^s(\Omega) \tag{6}$$

The following relation is a key step for our analysis.

$$\sum_{i=1}^{\infty} |\hat{f}_i|^2 \lambda_i^s = \int_\Omega f(x)\Delta^s f(x)dx = I_m^d(f) \tag{7}$$

$I_m^d(f)$ is a semi-norm as shown in relation (7). The null space is spanned by ϕ_1 since only $\lambda_1 = 0$. The Sobolev embedding theorem, together with relations (5), (6) and (7) imply that if we can bound semi-norm $I_m^d(f)$ for $2m > d$, function f will be continuous. Particularly, the semi-normed space is an RKHS. This provides a direct and simple solution to the degenerate problem in (Nadler et al., 2009).

2.3 SSL by Iterated Laplacian

In problem (2), $m = 1$. In order to obtain a continuous solution, we need $k \geq 0$ as in C^k. By the Sobolev embedding theorem, we need $2m > d$. Only $d = 1$ satisfies this inequality. This means problem (2) can only have continuous solutions for $d = 1$, i.e., either in $\mathbb{R}^1$ or on 1-dimensional submanifolds embedded in higher dimensions. This suggests that in order to find a continuous solution, we need $2m > d$. Therefore, we propose the following optimization problem for SSL

$$\min_f \sum_{x_i \in X_L} (f(x_i) - y_i)^2 + \mu I_{m,n}^d(f) \tag{8}$$

3 Limit Analysis of Iterated Laplacian Regularizer

In this section, by studying the limit of empirical iterated Laplacian regularizer $f^T L^m f$, we show that it is a good choice to implement the Sobolev semi-norm of order m. We first assume the density is uniform for simplicity, then discuss the nonuniform density case.

3.1 Uniform Density

Theorem 4. *Let Ω be a compact connected d-dimensional submanifold of $\mathbb{R}^N$ without boundary, $d \leq N$, data points $x_1, \cdots, x_n$ be drawn uniformly on Ω, and m be a positive integer. Assume $f(x) \in C^{2m}$, $Vol(\Omega) = 1$, then for $t_n = n^{-\frac{1}{d+2+\alpha}}$ where $\alpha > 0$, as $n \to \infty$ we have*

$$\frac{1}{n}(\frac{1}{nt_n^{d/2+1}})^m f^T L^m f \xrightarrow{p} \int_\Omega f(x)\Delta^m f(x)dx = I_m^d(f) \tag{9}$$

where t_n is the bandwidth of a Gaussian weight function, $n = l + u$, is the total number of data points.

Proof. We write $Lf(x)$ to mean the value at point x after applying discrete Laplacian to vector f. Based on the convergence of graph Laplacian from (Belkin and Niyogi, 2008, Theorem 3.1), for a fixed f and x, as $n \to \infty$, we have

$$\frac{1}{nt_n^{d/2+1}} Lf(x) \xrightarrow{p} \Delta f(x)$$

which implies for any $x \in \Omega$

$$f(x)\frac{1}{nt_n^{d/2+1}} Lf(x) \xrightarrow{p} f(x)\Delta f(x)$$

Since Lf can be seen as another vector, and $L(Lf)$ is just the application of operator L on Lf. When $f \in C^{2m}$, by law of large number, we have

$$\begin{aligned} &\tfrac{1}{n}(\tfrac{1}{nt_n^{d/2+1}})^m I_{m,n}^d(f) = \tfrac{1}{n}(\tfrac{1}{nt_n^{d/2+1}})^m f^T L^m f \\ &\xrightarrow{p} \int_\Omega f(x)\Delta^m f(x)dx = I_m^d(f) \end{aligned}$$

□

Therefore, when $2m > d$, theorem (1) guarantees that regularizer $I_m^d(f)$ is enough to restrict solutions to be continuous. By controlling m, we can even obtain C^k solutions with $k \geq 0$.

In the case of a uniform density, square loss with regularizer $I_m^d(f)$ on circles or other 1-dimensional domains without boundary is just thin place splines. However, for $m \geq 2$, As shown in theorem (4), regularizer $I_m^d(f)$ is different from regularizer $J_m^d(f)$ used in thin plate splines. Particularly, the null space of $I_m^d(f)$ for graph Laplacians is spanned only by $\phi_1(x)$, while the null of semi-norm $J_m^d(f)$ consists of polynomials of degree less than m. In thin plate splines, $2m > d$ is also required in order to obtain continuous solutions (Wahba, 1990).

When Ω has a smooth boundary, the limit of $I_{m,n}^d(f)$ is the same given proper boundary conditions. Since the technical details are beyond the scope of this paper, we leave a complete analysis for the future.

3.2 Nonuniform Density

For a nonuniform bounded density, we can similarly obtain the limit by the convergence of graph Laplacians with nonuniform density (Hein, 2005, Chapter 2). The only difference is that we have weighted Laplacians (Grigor'yan, 2006), instead of regular Laplacians. For example, consider the unnormalized graph Laplacian of the form $L = D - W$ for simplicity. We can rewrite $f^T L^m f$ as

$$f^T L^m f = \begin{cases} (L^{\frac{m}{2}} f)^T (L^{\frac{m}{2}} f), & \text{for even } m \\ \\ (L^{\frac{m-1}{2}} f)^T L (L^{\frac{m-1}{2}} f), & \text{for odd } m \end{cases}$$

On submanifolds without boundaries or submanifolds with boundaries with proper boundary conditions for f, as long as $Lf(x)$ is well defined for sufficiently smooth f, we can find $L^m f(x) = L[L^{(m-1)} f(x)]$ iteratively and find the limit accordingly. By plugging in the weighted Laplacian for the limit of L (Belkin and Niyogi, 2008), we can obtain a similar result as Theorem (4) using weighted Laplacians.

For weighted Laplacians, the highest order differential operator is always the regular Laplacian, therefore the limit of the iterated weighted Laplacian contains $I^d_m(f)$. This means for arbitrary bounded and smooth density, the iterated Laplacian regularizer can also restrict solution space to be an RKHS when $2m > d$.

4 Graph Laplacian, the Green's Function and Reproducing Kernel

There is a close relation between Green's functions and reproducing kernels, see e.g., (Wahba, 1990) and (Ramsay and Silverman, 1997, Chapter 20), which can be related by graph Laplacians. From partial differential equations (PDE's) point of view, solutions to PDE's can be written as a linear combination of Green's functions centered at labeled points. From RKHS view, the minimizer can also be written as a linear combination of kernels at labeled points.

4.1 Green's Functions

The Green's function $G(x, y)$ for Δ is a function of x such that

$$\Delta G(x, y) = \delta(x - y) \tag{10}$$

with proper boundary conditions. One way to obtain the Green's function is via eigenfunction expansion, see e.g., (Roach, 1982). When $\forall i, \lambda_i \neq 0$,

$$G(x, y) = \sum_{i=1}^{\infty} \frac{\phi_i(x)\phi_i(y)}{\lambda_i}$$

One potential difficulty is that for the Neumann eigenvalue, we have $\lambda_1 = 0$. This is the case for graph Laplacians and their limit operators. Then $G(x, y)$ is not well defined. In order to solve this problem, we define S_0 as a space spanned by ϕ_1, while S_1 as a space of all functions orthogonal to S_0. In S_1 our Neumann Green's function is

$$G(x, y) = \sum_{i=2}^{\infty} \frac{\phi_i(x)\phi_i(y)}{\lambda_i}$$

By eigenfunction expansion, we can also find the Green's function for Δ^m by using λ_i^m, since Δ^m is self-adjoint. Solutions to problem (1) then can be written as $f(x) = \sum_{x_i \in X_L} a_i G(x_i, x) + c$, where c is a constant. It can be shown when $d > 1$, we have $G(x, x) \to \infty$, and $\forall i, a_i \to 0$ in the limit of infinite unlabeled points, meaning the solution degenerates to constant c. This explains the degenerate solution problem in (Nadler et al., 2009) from another point of view.

4.2 Reproducing Property of the Green's Function

Consider semi-inner product $\langle u(x), v(x) \rangle_1 = \int_\Omega \nabla u(x) \cdot \nabla v(x) dx$. Note that in the subspace orthogonal to its null, this is an inner product, with induced norm $\|f(x)\|_1 = \int_\Omega \|\nabla f(x)\|^2 dx$. In S_1, the Green's function $G(x, t)$ has the reproducing property, as the following

$$\begin{aligned} & \langle f(\cdot), G(\cdot, x) \rangle_1 = \int_\Omega \nabla f(t) \cdot \nabla G(t, x) dt \\ = & \int_\Omega f(t) \cdot \Delta G(t, x) dt + \oint_{\partial\Omega} f(t) \nabla_{\mathbf{n}} G(t, x) dt \\ = & \int_\Omega f(t) \cdot \delta(t - x) dt = f(x) \end{aligned}$$

The boundary integral vanishes as a result of the empty boundary set or the Neumann boundary condition. When $G(x, x) \leq C < \infty$, the evaluation functionals are bounded by Cauchy-Schwarz inequality,

$$|f(x)| \leq \|G(\cdot, x)\|_1 \|f\|_1 = \sqrt{G(x, x)} \|f\|_1$$

This means $G(x, t)$ is a reproducing kernel for the RKHS with norm $\|\cdot\|_1$. However, when $G(x, x) \to \infty$, the corresponding normed space is not an RKHS. The boundary condition here makes sure the inner product and its induced norm are properly defined.

4.3 Finite Dimension Spaces

In finite data cases, we always have an RKHS since any finite dimensional Hilbert space is an RKHS, see e.g., (Berlinet and Thomas-Agnan, 2003). This means the discrete Green's function is the same as reproducing kernel in the subspace orthogonal to its null. Let the kernel matrix be K and discrete Green's function matrix be G, then for semi-norm $f^T L f$, the reproducing kernel in the subspace orthogonal to its null is the pseudoinverse[1] of matrix L, i.e., $K = L^+$ (Berlinet

[1] Notice that the exact kernel for the semi-norm includes another kernel in its null space.

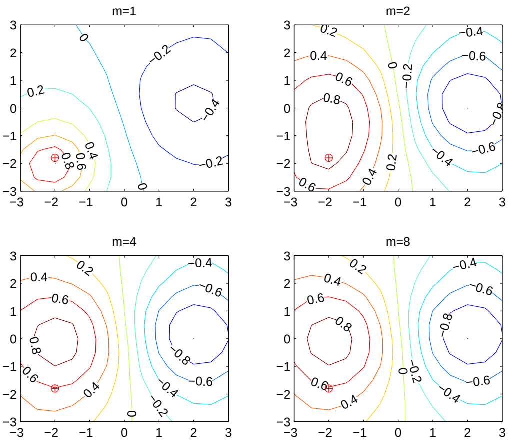

Figure 1: The Green's functions at (−2, −1.8) for a mixture of two Gaussians.

and Thomas-Agnan, 2003, Chapter 6). The discrete Green's function should satisfy $GL = I$ and $LG = I$, which implies that $G = L^+$. This is also true for symmetric semi-definite matrix L^m. We can write both discrete Green's function and reproducing kernel for L^m by eigenfunction expansion as

$$\begin{aligned} G_m(x,y) = \quad & K_m(x,y) = (L^m)^+(x,y) \\ = \quad & \textstyle\sum_{k=2}^{n} \frac{1}{(\lambda_k)^m}\phi_k(x)\phi_k(y) \end{aligned} \tag{11}$$

where λ_k and $\phi_k(x)$ are the k^{th} eigenvalue and eigenvector of L. $\phi_k(x)$ means the x element of vector ϕ_k. We can see that for a positive integer m, the smaller λ_k is (compared to λ_n), the larger $1/\lambda_k^m$ will be (compared to $1/\lambda_n^m$). This means $G_m(x,y)$ will become smoother and smoother as m increases, since for graph Laplacians, the smaller λ_k is, the smoother the associated eigenvector is, and λ_k are in increasing order.

In Figure (1), we show the discrete Green's function corresponding to L^m for two Gaussians with unit variance in $\mathbb{R}^2$. In order to bound density away from 0, we use a mixture of uniform with weight 0.2 and two well separated Gaussians with equal weight over $\mathbb{R}^2$, whose means are $(\pm 1.5, 0)$. We choose bandwidth t for Gaussian weight of order $n^{-1/(d+2)}$ (Belkin and Niyogi, 2008). As shown in Figure (1), when m increases, the Green's function or reproducing kernel "grows" from "spikes" to smooth functions[2]. From the contour plot, even the location of the kernel is not near the means of the Gaussians, when m increases, the kernel function recovers the true boundary of the two Gaussians. In fact, as long as the centers of kernel functions are in relatively high density regions of the Gaussians, a proper m value will produce a kernel function which changes little within those regions. This is exactly the basis for several graph Laplacian based SSL algorithms.

5 Experiments

In this section, we test the iterated graph Laplacian regularization method on a mixture of Gaussians and several real world datasets. We consider a transductive setting. The solution of problem (8) then becomes

$$\hat{f} = (S + \mu L^m)^+ SY \tag{12}$$

where Y is a column vector with $Y(i) = y_i$ for $x_i \in X_L$ and $Y(i) = 0$ for $x_i \in X_U$, $S = diag(1, \cdots, 1, 0, \cdots, 0)$ with the first l diagonal entries as 1 and the rest 0. Parameter t for the Gaussian weight and μ are common for most graph Laplacian based learning algorithms, while m controls the spectra transform. We test iterated Laplacian semi-norms with the following four empirical Laplacian matrices

$$\begin{aligned} L_u &= D - W \\ L_s &= D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \\ L_r &= D^{-1} L = I - D^{-1} W = I - P \\ L_g &= I - D_g^{-1} W_g \end{aligned}$$

where L_u is the unnormalized graph Laplacian, L_s the symmetric normalized Laplacian, L_r the random walk Laplacian, P the probability transition matrix on the same graph, and we call L_g the geometry Laplacian with $W_g = D^{-1} W D^{-1}$, D_g being the corresponding degree diagonal matrix $D_g(i,i) = \sum_j W_g(i,j)$. The limit operator of L_g is density independent (Coifman and Lafon, 2006).

One key difference between different versions of graph Laplacians is the density term hidden inside the limit operators when the density is not uniform. For instance, given a smooth function $f(x)$, $L_r f(x) \to \Delta_r f(x) = \Delta f(x) - \frac{2}{p(x)}\langle \nabla p(x), \nabla f(x)\rangle$ when $x \in \Omega/\partial\Omega$. However, for our regularizer $f^T L f$, there is another density term coming from the integration. For example, $L_g f(x) \to \Delta f(x)$, which has no density drifting term. This means this operator captures only the geometry of the submanifold. However, $f^T L_g f \to \int_\Omega f(x)\Delta f(x) p(x) dx$, which does have a density weight. Similarly, this problem also happens to inner product between eigenvectors of Laplacian. By using different parameters in normalizing the Laplacian, see (Hein, 2005; Coifman and Lafon, 2006), we can control the density term in the limit operator.

5.1 Mixture of Two Gaussians

We first test iterated Laplacians on a mixture of two Gaussians in $\mathbb{R}^{20}$, using normalized Laplacian L_s. We first generate two Gaussians in $\mathbb{R}^{20}$ with $\sigma_1 = \sigma_2 = 1$

[2] As we increase dimension d, the kernel will become a much sharper "spike" for $m = 1$.

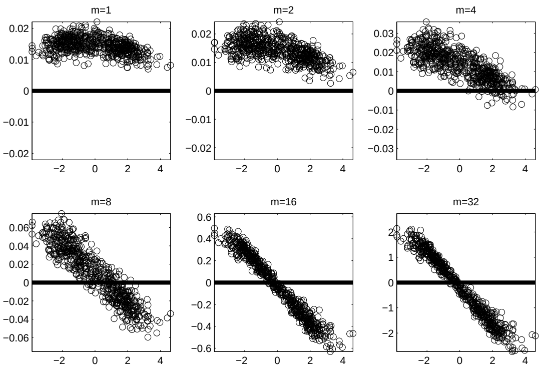

Figure 2: $\hat{f}(x)$ with different m on two Gaussians.

and $\mu_1 = \mu_2 = 0$, and then shift their means of the first dimension to $\mu_1 = -1.5$ and $\mu_2 = 1.5$. Complete graphs with Gaussian weight are used with $t = 20$. We use one labeled point for each class, the left Gaussian being $+1$ ($x_1 \leq 0$) while the right one being -1 ($x_1 > 0$), and $|X_U| = 600$ with balanced splits.

In Figure (2), we plot the estimated function by equation (12), with $\mu = 10^{-4}$. Points in $\mathbb{R}^{20}$ are projected to the first dimension as x axis, with means at ±1.5, and y axis shows the estimator $\hat{f}$. When $m = 1$, this is the ordinary Laplacian regularization method. We can easily see the two problems discussed in (Nadler et al., 2009), numerical instability and flat solutions. The whole solution shifts to the positive (or the negative side), which causes solutions to be unstable. Particularly in SSL, a small amount of random labeled examples can easily offset the whole solution to one side of x axis, even with a balanced class distribution. Moreover, differences between function values (y axis) of two classes are relatively small. Iterated Laplacians solve the two problems with a proper m. As m increases, we can see that $\text{sign}(\hat{f})$ recovers the true labels. First, it shifts solutions to the natural threshold zero as m increases, which makes solutions stable. Second, it increases the mean difference between the estimations on the two Gaussians.

5.2 Real World Data Sets

We test the iterated Laplacian method on high dimensional images, text, and gene data sets using symmetric normalized graph Laplacian L_s. Since we only focus on binary classification, we select 3vs8 and 4vs9 in MNIST. Standard binary text data includes aut-avn, real-sim, pcmac, ccat, gcat, and binary gene data sets are prostate cancer data (Gene-P) and breast cancer data (Gene-B) used in (Pochet et al., 2004)[3]. We divide each data set into three disjoint subsets, labeled set X_L, unlabeled set X_U and validation set X_V. $|X| = 1000$, $|X_V| = |X_L| = 50$ for MNIST 3vs8, 4vs9, pcmac, aut-avn, real-sim, ccat, gcat data sets, while Gene-P only has 136 data points and Gene-B has 97 points, so $|X_L| = |X_V| = 10$. We use simple settings for this experiment to emphasize the influence of iterated Laplacian semi-norm of order m. We use kNN graphs with $k = 20$, and $w_{ij} = 1$ if point x_i and x_j are neighbors, otherwise $w_{ij} = 0$, $\mu = 0.001$ [4]. The best m is selected among $\{1, 2, 4, 8, 16\}$ by validation on X_V. We run 100 random splits, and report the mean and standard deviation of classification errors and the average of the chosen m in Table (1). The iterated Laplacian method improves results across all datasets. Similar results hold for most reasonable parameters and graph settings. Results of different m correspond to the same random split.

Table 1: Classification errors % with std for $m = 1$ and adaptive m.

Data set	$m = 1$	IterLap	m
MNIST 3vs8	7.5 ± 1.5	6.2 ± 1.9	6.5
MNIST 4vs9	12.2± 2.9	8.4 ± 2.8	5.4
PCMAC	16.6± 2.4	12.5± 2.2	5.0
AUT-AVN	13.7± 2.6	11.0± 2.1	4.8
REAL-SIM	9.4 ± 3.3	6.2 ± 1.1	5.3
CCAT	24.0± 2.8	22.1± 3.1	5.6
GCAT	13.1± 1.5	12.7± 2.0	4.2
Gene-P	39.3± 9.1	29.7± 9.5	10.3
Gene-B	45.3± 7.3	41.9± 7.7	10.0

We also compare the results between the base case $m = 1$ with other fixed m, while the left parameters are chosen by validation. Results are shown in Table (2) for $m = 4$. Before m becomes too large, causing numerical issues, almost all $m \geq 2$ perform better than $m = 1$. We can see that for a fixed m, iterated Laplacian regularization also consistently outperforms base case $m = 1$. This is not surprising considering the influence of m to solutions presented in Figure (2). From the Sobolev embedding theorem, increasing m restricts solution space to be a smoother space, and from kernel point of view, increasing m corresponds to a better density adaptive kernel.

We also test the iterated Laplacian method on the binary classification bench mark in (Chapelle et al., 2006), using all four versions of empirical Laplacians. We add $\epsilon \|f\|^2$ to problem (8) to make the results comparable to (Chapelle et al., 2006, Chapter 11). The only difference is that we use an iterated Laplacian. Since the iterated Laplacian method can shift the threshold to zero, we did not use class mass normaliza-

[3] Prostate cancer dataset (Singh) and Breast cancer dataset (van't Veer).

[4] We test several other parameter settings, and the conclusions are similar.

Table 2: Classification errors % with std for $m = 1$ and $m = 4$.

Data set	$m = 1$	$m = 4$
MNIST 3vs8	7.5 ± 1.5	5.6 ± 1.3
MNIST 4vs9	12.2± 2.9	8.0 ± 2.6
pcmac	16.6± 2.4	11.5± 1.4
aut-avn	13.7± 2.6	10.1± 1.3
real-sim	9.4 ± 3.3	5.8 ± 0.8
ccat	24.0± 2.8	21.5± 2.8
gcat	13.1± 1.5	12.0± 1.3
Gene-P	39.3± 9.1	29.0± 8.9
Gene-B	45.3± 7.3	41.9± 9.2

tion (CMN). Following the setting of (Chapelle et al., 2006, Chapter 21), the best test error among four Laplacians is reported in Table(3), where "LapRLS" is Laplacian regularized least squares in (Chapelle et al., 2006, Chapter 12), and "Best" is the best result in (Chapelle et al., 2006, Chapter 21) among 13 different algorithms.

First, iterated Laplacians give large improvement compared to the base algorithm $m = 1$ (QC+CMN) on almost all datasets. The improvement is even larger compared to the base case without CMN. Second, results of iterated Laplacian method are also very competitive compared to other Laplacian related methods or even the best results, particularly on image and Gaussian datasets. Iterated Laplacian method performs relatively worse on Text, which is probably because complete graph is not a good choice for text data.

Notice that when we add $\epsilon\|f\|^2$ with $\epsilon > 0$ to the minimization objective of problem (8), the degenerate problem still exists since this normed space is $L^2(\Omega)$, which is already included in $H^1(\Omega)$. The solution becomes $\hat{f} = (S + \mu L^m + \epsilon I)^{-1}SY$. The effect of ϵ is to shift the spectra from λ_k to $\lambda_k + \epsilon$. From results in Table (3), we can also see that the density independent version of Laplacian L_g, in fact gives good results in practice. Although the limit of $L_g f(x)$ is independent of the underlying density, regularizer $f^T L_g^m f$ does depend on the density.

6 Discussion

There can be several variants of the iterated Laplacian regularizer besides the power of L. For example, we can add an RKHS norm to complete the semi-norm to a norm of the form $\|f\|^2_{\mathcal{H}_K} + \mu f^T L^m f$ as in manifold regularization (Belkin et al., 2006). Other possible variants are $f^T e^{-tL} f$, or $\sum_{k=1}^m \mu_k L^k$, such that $\mu_k \geq 0$ and $\sum \mu_k = 1$. Another interesting way is to use the linear combination of the pseudo-inverse of L^k as in multiple kernel learning (Argyriou et al., 2006).

Table 3: Classification errors % for SSL Benchmark.

	QC+CMN	LapRLS	Best	IterLap
		$\lvert X_L\rvert = 10$		
g241c	39.96	43.95	22.76	18.01 (L$_g$)
g241d	46.55	45.68	18.64	20.99 (L$_r$)
Digit1	9.80	5.44	5.44	6.54 (L$_g$)
USPS	13.61	18.99	13.61	13.10 (L$_s$)
BCI	50.36	48.97	46.90	46.71 (L$_u$)
Text	40.79	33.68	27.15	38.84 (L$_r$)
		$\lvert X_L\rvert = 100$		
g241c	22.05	24.36	13.49	14.82(L$_g$)
g241d	28.20	26.46	4.95	10.55(L$_g$)
Digit1	3.15	2.92	2.44	2.22(L$_g$)
USPS	6.36	4.68	4.68	3.96(L$_s$)
BCI	46.22	31.36	31.36	43.78(L$_s$)
Text	25.71	23.57	23.09	25.77(L$_r$)

Notice that the iterated Laplacian is not p-Laplacian as discussed in (Chapelle et al., 2006, Chapter 13), which is defined as

$$\Delta_p f = -\tfrac{1}{2}\mathrm{div}(\|\nabla f\|^{p-2}\nabla f)$$

which is a second order partial differential operator no matter what value p is.

Compared to the thin plate splines using $J^d_m(f)$ (Wahba, 1990), we can view iterated Laplacian regularization as a generalization of the thin plate splines from regular domains to unknown submanifolds, from a coordinate dependent Sobolev semi-norm defined by partial derivatives to a coordinate free iterated Laplacian semi-norm using Laplacians, from fixed data independent reproducing kernels to data dependent kernels. One key difference is the null space between the two methods. The null space of $f^T L^m f$ is spanned only by the first eigenvector, while in thin plate splines the null space is spanned by polynomials of high degrees, whose dimension increases fast as d and m increase.

Finally, the choice of m is important in practice. However, it is still unclear to us how to choose a good m other than validation.

Acknowledgements

We thank the anonymous reviewers for helping to improve the paper, and NSF for the support. XZ would like to thank Todd Dupont, Nathan Srebro and Boaz Nadler for valuable comments.

References

R. A. Adams. *Sobolev Spaces.* Academic Press, New York, 1975.

Andreas Argyriou, Mark Herbster, and Massimilano Pontil. Combining graph laplacians for semi–supervised learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 67–74, Cambridge, MA, 2006. MIT Press.

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, 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.

M. Belkin, P. Niyogi, and S. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

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

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.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning.* MIT Press, Cambridge, MA, 2006. URL `http://www.kyb.tuebingen.mpg.de/ssl-book`.

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.

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.

B. Nadler, N. Srebro, and X. 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.

N. Pochet, F. De Smet, J.Suykens, and B. De Moor. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, 20:3185–3195, 2004.

J. O. Ramsay and B. W. Silverman. *Functional data analysis.* Springer, 1997.

B. Daya Reddy. *Introductory Function Analysis, with Applications to Boundary Value Problems and Finite Elements.* Springer, 1997.

G. F. Roach. *Green's functions.* Cambridge University Press, 2nd edition, 1982.

A. Smola and I. Kondor. Kernels and regularization on graphs. In *Proc. 16th COLT*, pages 144–158, 2003.

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.

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, 2004.

X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin Madison, 2008.

X. Zhu, J. Lafferty, and Ghahramani, Z. Semi-supervised learning using gaussian fields and harmonic function. In *The Twentieth International Conference on Machine Learning*, 2003.