Harmonic Extension on Point Cloud

Abstract

In this paper, we consider the harmonic extension problem, which is widely used in many applications of machine learning. We find that the traditional method of graph Laplacian fails to produce a good approximation of the classical harmonic function. To tackle this problem, we propose a new method called the point integral method (PIM). We consider the harmonic extension problem from the point of view of solving Laplace equation on manifolds. The basic idea of the PIM method is to approximate the Laplace equation using an integral equation, which is easy to be discretized from points. Based on the integral equation, we explain the reason why the traditional graph Laplacian method (GLM) may fail to approximate the harmonic functions in the classical sense and propose a different approach which we call the point integral method (PIM). Theoretically, the PIM computes a harmonic function with convergence guarantees, and practically, it is also very easy to implement, which amount to solve a linear system. One important application of the harmonic extension in machine learning is semi-supervised learning. We run a popular semi-supervised learning algorithm by Zhu et al. (Zhu et al., 2003) over a couple of well-known datasets and compare the performance of the aforementioned approaches. Our experiments show the PIM performs the best. Finally, we apply PIM to an image recovery problem and show it outperforms GLM.

1. Introduction

In this paper, we consider the following harmonic extension problem. Let $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a set of points in $\mathbb{R}^d$ and B be a subset of X. Given a function $\mathbf{g}$ over B, let $C_{\mathbf{g}} = \{\mathbf{u} : X \to \mathbb{R} | \mathbf{u}_B = \mathbf{g}\}$ be the set of functions on X whose restriction to B coincides with $\mathbf{g}$. Denote $\mathbf{u}_i = \mathbf{u}(\mathbf{x}_i)$. The goal of the harmonic extension problem is to find the smoothest function in $C_{\mathbf{g}}$.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

A commonly used approach is based on graph Laplacian (Chung, 1997; Zhu et al., 2003). Let $w_{ij} = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{4t})$ be the Gaussian weight between $\mathbf{x}_i, \mathbf{x}_j$ for some parameter t. Consider the following quadratic energy functional over $C_{\mathbf{g}}$. For any $\mathbf{u} \in C_{\mathbf{g}}$ $E(\mathbf{u}) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(\mathbf{u}_i - \mathbf{u}_j)^2$. The small energy $E(\mathbf{u})$ means the function $\mathbf{u}$ takes similar values at nearby points, and the minimizer of this energy is considered as the smoothest function in $C_{\mathbf{g}}$. It is not difficult to see that the minimizer $\mathbf{u}$ satisfies that $\mathcal{L}\mathbf{u} = 0$ on the points in $X \setminus B$ and $\mathbf{u}_B = \mathbf{g}$. Here $\mathcal{L}$ is the weighted graph Laplacian given in matrix form as $\mathcal{L} = \frac{1}{t}(\mathcal{D} - \mathcal{W})$ where $\mathcal{W} = (w_{ij})$ is the weight matrix and $\mathcal{D} = \text{diag}(d_i)$ with $d_i = \sum_j w_{ij}$. We call $\mathbf{u}$ is discrete harmonic if $\mathcal{L}\mathbf{u} = 0$. The minimizer $\mathbf{u}$ can be computed by solving the linear system:

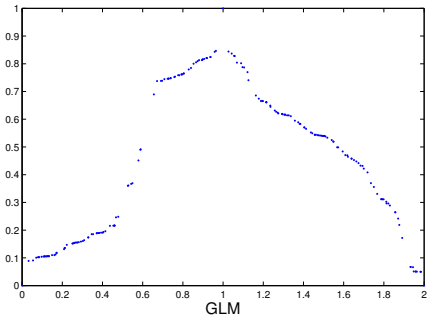
$$\begin{cases} \mathcal{L}(X \setminus B, X)\mathbf{u} = 0, \\ \mathbf{u}(\mathbf{x}_i) = \mathbf{g}(\mathbf{x}_i), \quad \forall \mathbf{x}_i \in B \end{cases} \tag{1}$$

where $\mathcal{L}(X \setminus B, X)$ is a submatrix of $\mathcal{L}$ by taking the rows corresponding to the subset $X \setminus B$. We call this approach of harmonic extension the graph Laplacian method (GLM). Note that the factor $\frac{1}{t}$ in $\mathcal{L}$ is immaterial in GLM but introduced to compare with other methods later.

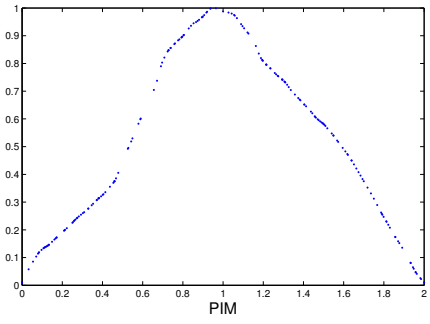
Consider the following simple example. Let X be the union of 198 randomly sampled points over the interval $(0, 2)$ and $B = \{0, 1, 2\}$. Set $\mathbf{g} = 0$ at $0, 2$ and $\mathbf{g} = 1$ at 1. We run the above graph Laplacian method over this example. Figure 1 (a) shows the resulting minimizer. It is well-known that the harmonic function over the interval $(0, 2)$ with the Dirichlet boundary $\mathbf{g}$, in the classical sense, is a piece linear function, i.e., $u(x) = x$ for $x \in (0, 1)$ and $u(x) = 2 - x$ for $x \in (1, 2)$; Clearly, the function computed by GLM does not approximate the harmonic function in the classical sense. In particular, the Dirichlet boundary has not been enforced properly, and in fact the obtained function is not even continuous near the boundary.

In this paper, we propose a new method which we call point integral method (PIM) to compute harmonic extension. Figure 1 (b) shows the harmonic function computed by PIM over the same data, which is a faithful approximation of the classical harmonic function. The point integral method is very simple and computes the harmonic extension by solving the following modified linear system:

$$\mathcal{L}\mathbf{u} + \mu\mathcal{W}(X, B)\mathbf{u}_B = \mu\mathcal{W}(X, B)\mathbf{g}, \tag{2}$$



(a)



(b)

Figure 1. One-dimensional examples

where $\mathcal{W}(X, B)$ is a submatrix of the weight matrix $\mathcal{W}$ by taking the columns corresponding to the subset B. Here the parameter μ is a fixed number whose choice will be described in Section 2. We consider the harmonic extension problem from the point of view of solving PDEs on manifolds, and derive the point integral method by approximating the Laplace equation using an *integral equation* which is then discretized using *points*. Note that the Dirichlet boundary may not be exactly enforced in PIM. Nevertheless, when the points X and B uniform randomly sample a submanifold and its boundary respectively in the iid fashion, the harmonic function computed by PIM is guaranteed to converge to the one in the classical sense.

We will give a brief derivation of the point integral method, and explain the reason that the graph Laplacian method fails to produce a faithful harmonic extension. For the details of the derivations and mathematical proofs, the interested readers are referred to the papers (Shi & Sun, c;b; Li et al.).

One important application of the harmonic extension in machine learning is semi-supervised learning (Zhu, 2005). We will perform the semi-supervised learning using the PIM over a couple of well-known data sets, and compare its performance to GLM as well as the closely related method by Zhou et al. (Zhou et al., 2003). The experimental results show that the PIM have the best performance.

Finally, we consider the image recovery problem where smoothness is a common regularity. Harmonic function is consider to be the most smooth funtion and wee use harmonic extension to recover the smooth function for image recovery. Again we compare the performaces of PIM and GLM in this application of image recovery.

1.1. Related work

The classical harmonic extension problem, also known as the Dirichlet problem for Laplace equation, has been studied by mathematicians for more than a century and has many applications in mathematics. The discrete harmonicity has also been extensively studied in the graph theory (Chung, 1997). For instance, it is closely related to random walk and electric networks on graphs (Doyle & Snell, 1984). In machine learning, the discrete harmonic extension and its variants have been used for semi-supervised learning (Zhu et al., 2003; Zhou et al., 2003).

Much of research has been done on the convergence of the graph Laplacian. When there is no boundary, the pointwise convergence of the graph Laplacian to the manifold Laplacian was shown in (Belkin & Niyogi, 2005; Lafon, 2004; Hein et al., 2005; Singer, 2006), and the spectral convergence of the graph Laplacian was shown in (Belkin & Niyogi, 2008). When there are boundaries, Singer and Wu (Singer & tieng Wu) and independently Shi and Sun (Shi & Sun, a) have shown that the spectra of the graph Laplacian converge to that of manifold Laplacian with Neumann boundary.

The discrete harmonic extension problem was studied in the community of numerical PDEs and its convergence result is well-known if the Laplacian matrix is derived based on finite difference provided that the points lie on a regular grid, or based on finite element provided that the points are the vertices of a well-shaped mesh tessellating the domain. However, both assumptions on the points are difficult to be satisfied in the applications of machine learning. Du et al. (Du et al., 2012) considered the nonlocal diffusion problems which is modeled by a similar integral equation. They observed that the regularity of the boundary condition can not infer the regularity of the harmonic extension and thus proposed to thicken the boundary and employed volume constraint.

2. Point Integral Method

Harmonic Extension: We consider the harmonic extension in the continuous form, as shown in Figure 2. Assume $\mathcal{M}$ is a submanifold embedded in $\mathbb{R}^d$. Consider a function $u(\mathbf{x})$ defined on $\mathcal{M}$ and $u(\mathbf{x})$ is known in some regions $\Omega_1 \cup \cdots \cup \Omega_k \subset \mathcal{M}$. Now, we want to extend the function $u(\mathbf{x})$ from $\Omega_1 \cup \cdots \cup \Omega_k$ to the entire manifold $\mathcal{M}$. If the

manifold $\mathcal{M}$ has boundary, $\partial\mathcal{M}$, we assume $\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = 0$ at $\mathcal{M}$. Then, the harmonic extension problem shown in Figure 2 can be modelled as the following Laplace equation with the mixed boundary condition:

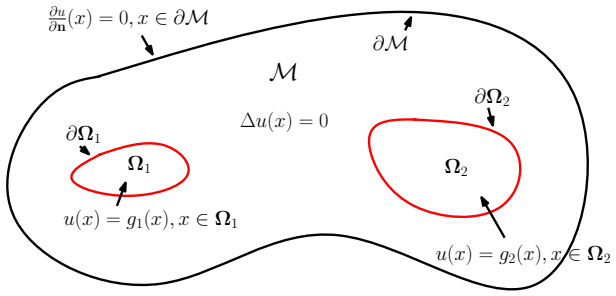


Figure 2. Sketch of the manifold.

$$\begin{cases} -\Delta_{\mathcal{M}} u(\mathbf{x}) = 0, & \mathbf{x} \in \mathcal{M}, \\ u(\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \partial\mathcal{M}_D, \\ \frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = 0, & \mathbf{x} \in \partial\mathcal{M}_N. \end{cases} \tag{3}$$

In the aforementioned harmonic extension problem, one can think of $\partial\mathcal{M}_D$ as the boundary of $\Omega_1 \cup \cdots \cup \Omega_k$, and $\partial\mathcal{M}_N$ as the actual boundary of $\mathcal{M}$.

Integral Equation: We observe that the Laplace equation $\Delta\mathcal{M} = 0$ is closely related to the following integral equation.

$$\frac{1}{t}\int_{\mathcal{M}} (u(\mathbf{x}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} - 2\int_{\partial\mathcal{M}} \frac{\partial u(\mathbf{y})}{\partial \mathbf{n}} w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}} = 0, \tag{4}$$

where $w_t(\mathbf{x},\mathbf{y}) = \exp(-\frac{|\mathbf{x}-\mathbf{y}|^2}{4t})$.

Next, we give a brief derivation of the integral equation (4) in the Euclidean space. We assume $\mathcal{M}$ is an open set on $\mathbb{R}^d$. For a general submanifold, the derivation follows from the same idea but is technically more involved. The interested readers are referred to (Shi & Sun, c). Thinking of $w_t(\mathbf{x},\mathbf{y})$ as test functions, by integral by parts, we have

$$\begin{aligned} &\int_{\mathcal{M}} \Delta u w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ =&-\int_{\mathcal{M}} \nabla u\cdot\nabla w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} + \int_{\partial\mathcal{M}} \frac{\partial u}{\partial \mathbf{n}} w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}} \\ =&\frac{1}{2t}\int_{\mathcal{M}} (\mathbf{y}-\mathbf{x})\cdot\nabla u(\mathbf{y}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ &+\int_{\partial\mathcal{M}} \frac{\partial u}{\partial \mathbf{n}} w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}}. \end{aligned} \tag{5}$$

The Taylor expansion of the function u tells us that

$$\begin{aligned} &u(\mathbf{y}) - u(\mathbf{x}) \\ =&(\mathbf{y}-\mathbf{x})\cdot\nabla u(\mathbf{y}) - \frac{1}{2}(\mathbf{y}-\mathbf{x})^T\mathbf{H}_u(\mathbf{y})(\mathbf{y}-\mathbf{x}) \\ &+O(\|\mathbf{y}-\mathbf{x}\|^3), \end{aligned}$$

where $\mathbf{H}_u(\mathbf{y})$ is the Hessian matrix of u at $\mathbf{y}$. Note that $\int_{\mathcal{M}} \|\mathbf{y}-\mathbf{x}\|^n w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} = O(t^{n/2})$. We only need to estimate the following term.

$$\begin{aligned} &\frac{1}{4t}\int_{\mathcal{M}} (\mathbf{y}-\mathbf{x})^T\mathbf{H}_u(\mathbf{y})(\mathbf{y}-\mathbf{x}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ =&\frac{1}{4t}\int_{\mathcal{M}} (\mathbf{y}_i-\mathbf{x}_i)(\mathbf{y}_j-\mathbf{x}_j)\partial_{ij}u(\mathbf{y}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ =&-\frac{1}{2}\int_{\mathcal{M}} (\mathbf{y}_i-\mathbf{x}_i)\partial_{ij}u(\mathbf{y})\partial_j\left(w_t(\mathbf{x},\mathbf{y})\right)\mathrm{d}\mathbf{y} \\ =&\frac{1}{2}\int_{\mathcal{M}} \partial_j(\mathbf{y}_i-\mathbf{x}_i)\partial_{ij}u(\mathbf{y}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ &+\frac{1}{2}\int_{\mathcal{M}} (\mathbf{y}_i-\mathbf{x}_i)\partial_{ijj}u(\mathbf{y}) w_t(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{y} \\ &-\frac{1}{2}\int_{\partial\mathcal{M}} (\mathbf{y}_i-\mathbf{x}_i)\mathbf{n}_j\partial_{ij}u(\mathbf{y}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}} \\ =&\frac{1}{2}\int_{\mathcal{M}} \Delta u(\mathbf{y}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ &-\frac{1}{2}\int_{\partial\mathcal{M}} (\mathbf{y}_i-\mathbf{x}_i)\mathbf{n}_j\partial_{ij}u(\mathbf{y}) w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}} + O(t^{1/2}). \end{aligned} \tag{6}$$

Now consider the second summand in the last line. Although its $L_\infty(\mathcal{M})$ norm is of the constant order, its $L^2(\mathcal{M})$ norm is of the order $O(t^{1/2})$ due to the fast decay of $w_t(\mathbf{x},\mathbf{y})$. Therefore, we could obtain Theorem 2.1, which follows from the equations (5) and (6).

Theorem 2.1 *If $u \in C^3(\mathcal{M})$ be a harmonic function on $\mathcal{M}$, i.e., $\Delta_{\mathcal{M}} u = 0$, then we have for any $\mathbf{x} \in \mathcal{M}$,*

$$\|r(u)\|_{L^2(\mathcal{M})} = O(t^{1/4}). \tag{7}$$

where

$$r(u) = \frac{1}{t}\int_{\mathcal{M}} (u(\mathbf{x}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} - 2\int_{\partial\mathcal{M}} \frac{\partial u(\mathbf{y})}{\partial \mathbf{n}} w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}}$$

The detailed proof can be found in (Shi & Sun, c).

Dirichlet Boundary: Using the integral equation (4) and the boundary condition of (3), we know the Laplace equation with the mixed boundary condition can be approximated by following integral equation:

$$\frac{1}{t}\int_{\mathcal{M}} (u(\mathbf{x}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} - 2\int_{\partial\mathcal{M}_D} \frac{\partial u(\mathbf{y})}{\partial \mathbf{n}} w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}} = 0, \tag{8}$$

However, on $\partial\mathcal{M}_D$, $\frac{\partial u}{\partial \mathbf{n}}$ is not known. To solve this issue, we use the Robin boundary condition to approximate the original Dirichlet boundary condition on $\partial\mathcal{M}_D$. Then, we consider the following Robin/Neumann mixed boundary problem.

$$\begin{cases} -\Delta_{\mathcal{M}} u(\mathbf{x}) = 0, & \mathbf{x} \in \mathcal{M}, \\ u(\mathbf{x}) + \beta \frac{\partial u(\mathbf{x})}{\partial \mathbf{n}} = g(\mathbf{x}), & \mathbf{x} \in \partial\mathcal{M}_D, \\ \frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = 0, & \mathbf{x} \in \partial\mathcal{M}_N. \end{cases} \tag{9}$$

where $\beta > 0$ is a parameter. It is easy to prove that the solution of the above Robin/Neumann problem (9) converges to the solution of the Dirichlet/Neumann problem (3) as β goes to 0.

By substituting the Robin boundary $\frac{\partial u(\mathbf{x})}{\partial \mathbf{n}} = \frac{1}{\beta}(g(\mathbf{x}) - u(\mathbf{x}))$ in the integral equation (8), we get an integral equation to solve the Robin/Neumann problem.

$$\begin{aligned} &\frac{1}{t}\int_{\mathcal{M}} (u(\mathbf{x}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} \\ &\quad - \frac{2}{\beta}\int_{\partial\mathcal{M}_D} (g(\mathbf{y}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}} = 0, \end{aligned} \tag{10}$$

When $\beta > 0$ is small enough, this integral equation also gives a good approximation to the original harmonic extension problem (3).

2.1. Discretization

Denote

$$L_t u(\mathbf{x}) = \frac{1}{t}\int_{\mathcal{M}} (u(\mathbf{x}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y}, \tag{11}$$

$$I_t u(\mathbf{x}) = \int_{\partial\mathcal{M}_D} (g(\mathbf{y}) - u(\mathbf{y}))w_t(\mathbf{x},\mathbf{y})\mathrm{d}\tau_{\mathbf{y}}. \tag{12}$$

Assume that the point set $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ samples the submanifold $\mathcal{M}$ and it is uniformly distributed, then $L_t u$ is discretized and well approximated by $\mathcal{L}\mathbf{u}$ up to the volume weight $\frac{|\mathcal{M}|}{n}$ where $\mathbf{u} = (u(\mathbf{x}_1), \cdots, u(\mathbf{x}_n))$. Recall that $\mathcal{L}$ is the weighted graph Laplacian given in matrix form as $\mathcal{L} = \frac{1}{t}(\mathcal{D} - \mathcal{W})$ where $\mathcal{W} = (w_{ij})$ is the weight matrix and $\mathcal{D} = \mathrm{diag}(d_i)$ with $d_i = \sum_j w_{ij}$. In this paper, we use the Gaussian weight, $w_{ij} = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{4t})$.

The boundary term $I_t u(\mathbf{x})$ is actually corresponding to the subset $B = \{\mathbf{b}_1, \cdots, \mathbf{b}_m\} \subset X$ where the values of function u are given. From the continuous point of view, for each point $\mathbf{b}_i \in B$, in a small area around it, the value of u is given. In this sense, each $\mathbf{b}_i$ actually stands for one part of the boundary $\partial\mathcal{M}_D$. Based on the above discussion, the boundary term $I_t u(\mathbf{x})$ can be discretized as $\sum_{\mathbf{b}_i \in B} w_t(\mathbf{x}, \mathbf{b}_i)(g(\mathbf{b}_i) - u(\mathbf{b}_i))$ up to the surface area weight $\frac{|\partial\mathcal{M}_D|}{m}$.

Therefore, the harmonic extension problem (3) can be numerically solved by the following linear system

$$\mathcal{L}\mathbf{u} - \frac{2}{\beta}\frac{n|\partial\mathcal{M}_D|}{m|\mathcal{M}|}\mathcal{W}(X,B)(\mathbf{g} - \mathbf{u}_B) = 0 \tag{13}$$

where $\mathcal{W}(X,B)$ is a submatrix of the weight matrix $\mathcal{W}$ by taking the columns corresponding to the subset B, and $\mathbf{g} = (g(\mathbf{b}_1), \cdots, g(\mathbf{b}_m))$ and $\mathbf{u}_B = (u(\mathbf{b}_1), \cdots, u(\mathbf{b}_m))$. This is the same as the linear system (2) if set $\mu = \frac{2}{\beta}\frac{n|\partial\mathcal{M}_D|}{m|\mathcal{M}|}$. In the examples shown in the paper, we take $\beta = 10^{-4}\frac{|\partial\mathcal{M}_D|}{|\mathcal{M}|}$, and thus $\mu = 10^4\frac{n}{m}$.

When the points X and B uniform randomly sample a submanifold and its boundary respectively in the iid fashion, the harmonic function computed by PIM is guaranteed to converge to the one in the classical sense. For the details of the theoretical analysis, we refer to (Shi & Sun, c;b).

Remark 2.1 *Based on the discussion in this section, we can see clearly the reason that the traditional graph Laplacian may fail to approximate the classic harmonic functions. The reason is that in the graph Laplacian approach, the boundary term is dropped. However, this boundary term is not small. Without this term, the boundary condition may not be enforced correctly. This effect has been shown in Figure 1 and more evidence will be given in the example section.*

3. Semi-supervised Learning

In this section, we briefly describe the algorithm of semi-supervised learning based on the harmonic extension proposed by Zhu et al. (Zhu et al., 2003). We plug into the algorithm the aforementioned approach for harmonic extension, and apply them to several well-known data sets, and compare their performance.

Assume we are given a point set $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_m, \mathbf{x}_{m+1}, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^d$, and a label set $\{1, 2, \cdots, l\}$, and the label assignment on the first m points $L : \{\mathbf{x}_1, \cdots, \mathbf{x}_m\} \to \{1, 2, \cdots, l\}$. In a typical setting, m is much smaller than n. The purpose of the semi-supervised learning is to extend the label assignment L to the entire X, namely, infer the labels for the unlabeled points.

Think of the label points as the boundary $B = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$. For the label i, we set up the Dirichlet boundary $\mathbf{g}^i$ as follows. If a point $\mathbf{x}_j \in B$ is labelled as i, set $\mathbf{g}^i(\mathbf{x}_j) = 1$, and otherwise set $\mathbf{g}^i(\mathbf{x}_j) = 0$. Then we compute the harmonic extension $\mathbf{u}^i$ of $\mathbf{g}^i$ using the aforementioned approaches. In this way, we obtain a set of l harmonic functions $\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^l$. We label $\mathbf{x}_j$ using k where $k = \arg\max_{i \le l} \mathbf{u}^i(\mathbf{x}_j)$. The algorithm is summarized in Algorithm 1. Note that this algorithm is slightly

Algorithm 1 Semi-Supervised Learning

Require: A point set $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_m, \mathbf{x}_{m+1}, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^d$ and a partial label assignment $L : \{\mathbf{x}_1, \cdots, \mathbf{x}_m\} \to \{1, 2, \cdots, l\}$

Ensure: A complete label assignment $L : X \to \{1, 2, \cdots, l\}$

```
for i = 1 : l do
    for j = 1 : m do
        Set g^i(x_j) = 1 if L(x_j) = i, and otherwise set g^i(x_j) = 0.
    end for
    Compute the harmonic extension u^i of g^i.
end for
for j = m + 1 : n do
    L(x_j) = k where k = arg max_{i≤l} u^i(x_j).
end for
```

different from the original algorithm by Zhu et al. (Zhu et al., 2003) where only one harmonic extension was computed by setting $\mathbf{g}^i(\mathbf{x}_j) = k$ if $\mathbf{x}_j$ has a label k.

3.1. Experiments

We now apply the above semi-supervised learning algorithm to a couple of well-known data sets: MNIST and 20 Newsgroups. We do not claim the state of the art performance on these datasets. The purpose of these experiments is to compare the performance of different approaches of harmonic extension. We also compare to the closely related method of local and global consistency by Zhou et al. (Zhou et al., 2003). Besides these two examples, we also apply PIM on one example related with image recovery.

MNIST : In this experiment, we use the MNIST of dataset of handwritten digits (Burges et al.), which contains $60k$ 28×28 gray scale digit images with labels. We view digits $0 \sim 9$ as ten classes. Each digit can be seen as a point in a common 784-dimensional Euclidean space. We randomly choose 16k images. Specifically, there are 1606, 1808, 1555, 1663, 1552, 1416, 1590, 1692, 1521 and 1597 digits in $0 \sim 9$ class respectively.

To set the parameter t, we build a graph by connecting a point x_i to its 10 nearest neighbors under the standard Euclidean distance. We compute the average of the distances for x_i to its neighbors on the graph, denoted h_i. Let h be the average of h_i's over all points and set $t = h^2$. The distance $|\mathbf{x}_i - \mathbf{x}_j|$ is computed as the graph distance between x_i and x_j. In the method of local and global consistency, we follow the paper (Zhou et al., 2003) and set the width of the RBF kernel to be 0.3 and the parameter α in the iteration process to be 0.3.

For a particular trial, we choose k $(k = 1, 2, \cdots, 10)$ images randomly from each class to assemble the labelled set B and assume all the other images are unlabelled. For each fixed k, we do 100 trials. The error bar of the tests is presented in Figure 3 (a). It is quite clear that the PIM has the best performance when there are more than 5 labelled points in each class, and the GLM has the worst performance.

Newsgroup: In this experiment, we use the 20-newsgroups dataset, which is a classic dataset in text classification. We only choose the articles from topic *rec* containing four classes from the version 20-news-18828. We use Rainbow (version:20020213) to pre-process the dataset and finally vectorize them. The following command-line options are required[1]: (1)- *-skip-header*: to avoid lexing headers; (2)- *-use-stemming*: to modify lexed words with the 'Porter' stemmer; (3)- *-use-stoplist*: to toss lexed words that appear in the SMART stoplist; (4)- *-prune-vocab-by-doc-count=5*: to remove words that occur in 5 or fewer documents; Then, we use TF-IDF algorithm to normalize the word count matrix. Finally, we obtain 3970 documents (990 from rec.autos, 994 from rec.motorcycles, 994 from rec.sport.baseball and 999 from rec.sport.hockey) and a list of 8014 words. Each document will be treated as a point in a 8014-dimensional space.

To deal with text-kind data, we define a new distance introduced by Zhu et al. (Zhu et al., 2003): the distance between x_i and x_j is $d(x_i, x_j) = 1 - \cos\alpha$, where α is the angle between x_i and x_j in Euclidean space. Under this new distance, we ran the same experiment with the same parameter as we process the above MNIST dataset. The error bar of the tests for 20-newsgroups is presented in Figure 3 (b). A similar pattern result is observed, namely the PIM has the best performance when there are more than 2 labelled points in each class, and the GLM has the worst performance.

4. Image Recovery

In this example, we consider an image recovery problem. The original image is shown in Figure 4(a) which has 512×512 pixels. Then, we subsample the image and only retain 1% of the pixels. The positions of the retained pixels are selected at random. The subsampled image is shown in Figure 4(b). Now, we want to recover the original image from the subsampled image. This is a classical problem in image processing which has been studied extensively. Here, we only use this example to demonstrate the difference between PIM method and the Graph Laplacian approach, rather than presenting an image recovery method.

First, we construct a point cloud from the original image, denoted by f, by using so called patch approach which

[1] all the following options are offered by Rainbow

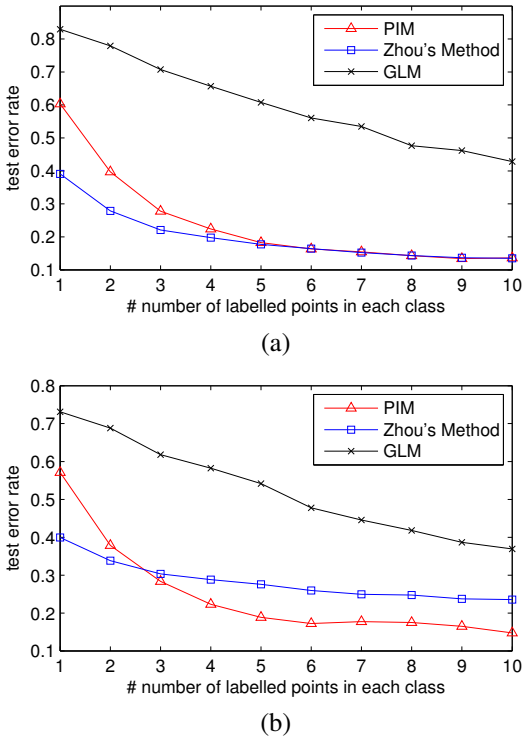


Figure 3. (a) the error rates of digit recognition with a 16000-size subset of MNIST dataset; (b) the error rates of text classification with 20-newsgroups.rec(a 8014-dimensional space with 3970 data points).

is widely used in image processing. For each pixel x_i in the image f, we extract a patch around it of size 3×3 which is denoted as $p_{x_i}(f)$. Here $i = 1, \cdots, 512^2$. For the pixels on the boundary, the patch is obtained by extending the image symmetrically. Then, we can get 512^2 patches and each patch is 3×3. These patches consist of a point cloud in $\mathbb{R}^9$. Denote this point cloud as $P = \{p_{x_i}(f) : i = 1, \cdots, 512^2\}$. And function u on P is defined as $u(p_{x_i}(f)) = f(x_i)$, $f(x_i)$ is the value of image f at pixel x_i. Using this definition, at some patches which around the retained pixels, the value of u is known. The collection of these patches is denoted as S which is a subset of P.

Inspired by the nonlocal method in image processing (Buades et al., 2005; Gilboa & Osher, 2008), the function u should be a smooth function over P. Notice that u on $S \subset P$ is known, one natural approach to recover u is harmonic extension, i.e. solving following linear system

$$\mathcal{L}_P \mathbf{v} - \mu \mathcal{W}(P, S)(\mathbf{u}_S - \mathbf{v}_S) = 0 \tag{14}$$

Here $\mathcal{L}_P$ is the weighted graph Laplacian over the point cloud P which is given in matrix form as $\mathcal{L}_P = \frac{1}{t}(\mathcal{D} - \mathcal{W})$ where $\mathcal{W} = (w_{ij})$ is the weight matrix and $\mathcal{D} = \mathrm{diag}(d_i)$ with $d_i = \sum_j w_{ij}$. $\mathcal{W}(P, S)$ is a submatrix of the weight matrix $\mathcal{W}$ by taking the columns corresponding to the subset S. $\mathbf{u}_S$ is a vector consists of the value of u on S. $\mathbf{v}_S$ is a vector by confining $\mathbf{v}$ on S.

We also compute the solution given by graph Laplacian, i.e., solving following linear system.

$$\begin{cases} \mathcal{L}_P(P \setminus S, P)\mathbf{v} = 0, \\ \mathbf{v}(\mathbf{x}_i) = u(\mathbf{x}_i), \quad \forall \mathbf{x}_i \in S \end{cases} \tag{15}$$

where $\mathcal{L}_P(P \setminus S, P)$ is a submatrix of $\mathcal{L}_P$ by taking the rows corresponding to the subset $P \setminus S$.

We remark that in this example the point cloud P is constructed using the original image shown in Figure 4(a). So this is not an full image recovery method since the original image is used.

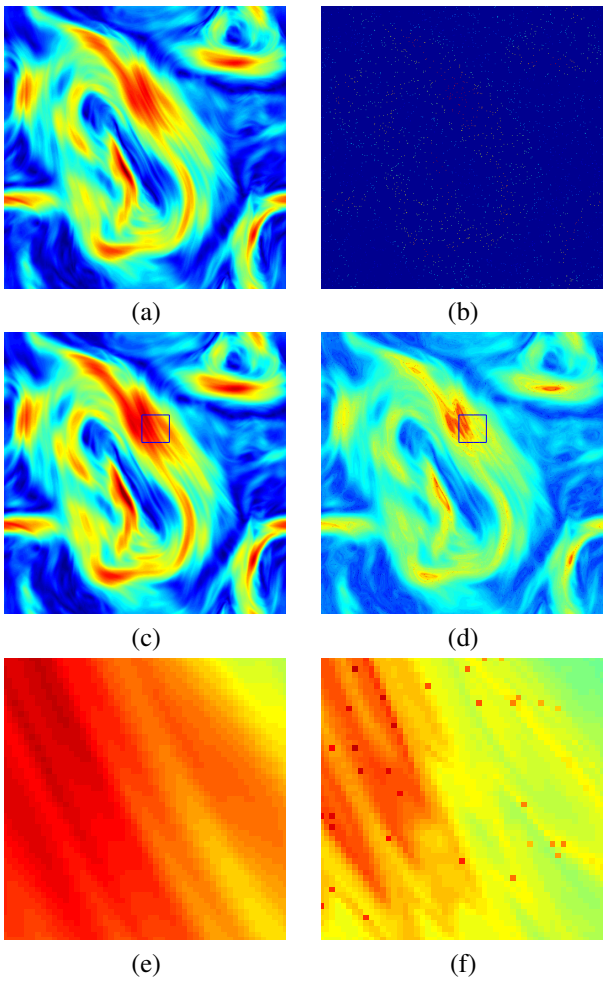


Figure 4. (a) original image; (b) subsampled image (1% pixels are retained); (c) recovered image by PIM; (d) recovered image by GL; (e) zoom in image of (c); (f) zoom in image of (d).

The solution of PIM (14) is given in Figure 4(c) and the solution of GLM (15) is given in Figure 4(d). Obviously, the result given by PIM is much better. To get a closer look at

of the recovery, Figure 4(e), (f) show the zoom in image enclosed by the boxes in Figure 4(c), (d) respectively. In Figure 4(f), there are many pixels which are not consistent with their neighbors. And it is easy to check that these pixels are actually the retained pixels. This phenomenon suggests that in GLM, (15), the values at the retained pixels are not spreaded to their neighbours properly. The reason is that in GLM a non-negligible boundary term is dropped as we pointed in this paper. On the contrary, in PIM, the boundary term is retained and the resultant recovery is much better and smoother as shown in 4(c)(e).

5. Conclusion

We have presented a novel approach, point integral method, for solving the harmonic extension problem. We have also compared its performance with that of graph Laplacian in the application of semi-supervised learning and image recovery. In the future, we will test this method on more datasets and find different applications of harmonic extension.

References

Belkin, Mikhail and Niyogi, Partha. Towards a theoretical foundation for laplacian-based manifold methods. In *COLT*, pp. 486–500, 2005.

Belkin, Mikhail and Niyogi, Partha. Convergence of laplacian eigenmaps. *preprint, short version NIPS 2008*, 2008.

Buades, A., Coll, B., and Morel, J.-M. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, 4:490–530, 2005.

Burges, Christopher J.C., LeCun, Yann, and Cortes, Corinna. Mnist database.

Chung, F. R. K. *Spectral Graph Theory*. American Mathematical Society, 1997.

Doyle, Peter G. and Snell, J. Laurie. *Random Walks and Electric Networks*. Mathematical Association of America, Washington, DC, 1984.

Du, Qiang, Gunzburger, Max, Lehoucq, Richard B., and Zhou, Kun. Analysis and approximation of nonlocal diffusion problems with volume constraints. *SIAM Review*, 54:667–696, 2012.

Gilboa, G. and Osher, S. Nonlocal operators with applications to image processing. *Multiscale Model. Simul.*, 7: 1005–1028, 2008.

Hein, Matthias, Audibert, Jean-Yves, and von Luxburg, Ulrike. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In *Proceedings of the 18th Annual Conference on Learning Theory*, COLT'05, pp. 470–485, 2005. ISBN 3-540-26556-2, 978-3-540-26556-6.

Lafon, S. *Diffusion Maps and Geodesic Harmonics*. PhD thesis, 2004.

Li, Zhen, Shi, Zuoqiang, and Sun, Jian. Point integral method for solving poisson-type equations on manifolds from point clouds with convergence guarantees. *arXiv:1409.2623*.

Shi, Zuoqiang and Sun, Jian. Convergence of laplacian spectra from point clouds. *arXiv:1506.01788*, a.

Shi, Zuoqiang and Sun, Jian. Convergence of the point integral method for the poisson equation on manifolds ii: the dirichlet boundary. *arXiv:1312.4424*, b.

Shi, Zuoqiang and Sun, Jian. Convergence of the point integral method for the poisson equation on manifolds i: the neumann boundary. *arXiv:1403.2141*, c.

Singer, A. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21:128–134, 2006.

Singer, Amit and tieng Wu, Hau. Spectral convergence of the connection laplacian from random samples. *arXiv:1306.1587*.

Zhou, Dengyong, Bousquet, Olivier, Lal, Thomas Navin, Weston, Jason, and Schölkopf, Bernhard. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pp. 321–328, 2003.

Zhu, Xiaojin. *Semi-supervised Learning with Graphs*. PhD thesis, Pittsburgh, PA, USA, 2005. AAI3179046.

Zhu, Xiaojin, Ghahramani, Zoubin, and Lafferty, John D. Semi-supervised learning using gaussian fields and harmonic functions. In *Machine Learning, Proceedings of the Twentieth International Conference ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 912–919, 2003.