An Analysis of the Convergence of Graph Laplacians

Daniel Ting DTING@STAT.BERKELEY.EDU
Department of Statistics, University of California, Berkeley, CA 94720 USA

Ling Huang LING.HUANG@INTEL.COM
Intel Labs Berkeley, 2150 Shattuck Avenue, Berkeley, CA 94704 USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of EECS and Department of Statistics, University of California, Berkeley, CA 94720 USA

Abstract

Existing approaches to analyzing the asymptotics of graph Laplacians typically assume a well-behaved kernel function with smoothness assumptions. We remove the smoothness assumption and generalize the analysis of graph Laplacians to include previously unstudied graphs including kNN graphs. We also introduce a kernel-free framework to analyze graph constructions with shrinking neighborhoods in general and apply it to analyze locally linear embedding (LLE). We also describe how, for a given limit operator, desirable properties such as a convergent spectrum and sparseness can be achieved by choosing the appropriate graph construction.

1. Introduction

Graph Laplacians have become a core technology throughout machine learning. In particular, they have appeared in clustering (Kannan et al., 2004; von Luxburg et al., 2008), dimensionality reduction (Belkin & Niyogi, 2003; Nadler et al., 2006), and semi-supervised learning (Zhu, 2008).

While graph Laplacians are but one member of a broad class of methods that use local neighborhood graphs to model high-dimensional data lying on a low-dimensional manifold, they are distinguished by their appealing mathematical properties, notably: (1) the graph Laplacian is the infinitesimal generator for a random walk on the graph, and (2) it is a discrete approximation to a weighted Laplace-Beltrami operator on a manifold, an operator which has numerous geometric properties and induces a smoothness functional. These mathematical properties have served as a foundation for the development of a growing theoretical literature that has analyzed learning procedures based on the graph Laplacian. To review briefly, Bousquet et al. (2003) proved an early result for the convergence of the unnormalized graph Laplacian to a smoothness functional that depends on the squared density p^2. Belkin & Niyogi (2005) demonstrated the pointwise convergence of the empirical unnormalized Laplacian to the Laplace-Beltrami operator on a compact manifold with uniform density. Lafon (2004) and Nadler et al. (2006) established a connection between graph Laplacians and the infinitesimal generator of a diffusion process. They further showed that one may use the degree operator to control the effect of the density. Hein et al. (2005) combined and generalized these results for weak and pointwise (strong) convergence under weaker assumptions, as well as providing rates for the unnormalized, normalized, and random walk Laplacians. They also make explicit the connections to the weighted Laplace-Beltrami operator. Singer (2006) obtained improved convergence rates for a uniform density. Giné & Koltchinskii (2005) established a uniform convergence result and functional central limit theorem to extend the pointwise convergence results. von Luxburg et al. (2008) and Belkin & Niyogi (2006) presented convergence results for the eigenvectors of graph Laplacians in the fixed and shrinking bandwidth cases respectively.

Although this burgeoning literature has provided many useful insights, several gaps remain between theory and practice. Most notably, in constructing the neighborhood graphs underlying the graph Laplacian, several choices must be made, including the choice of algorithm for constructing the graph, with k-nearest-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

neighbor (kNN) and kernel functions providing the main alternatives, as well as the choice of parameters (k, kernel bandwidth, normalization weights). These choices can lead to the graph Laplacian generating fundamentally different random walks and approximating different weighted Laplace-Beltrami operators. The existing theory has focused on one specific choice in which graphs are generated with smooth kernels with shrinking bandwidths. But a variety of other choices are often made in practice, including kNN graphs, r-neighborhood graphs, and the "self-tuning" graphs of Zelnik-Manor & Perona (2004). Surprisingly, few of the existing convergence results apply to these choices (see Maier et al. (2008) for an exception).

This paper provides a general theoretical framework for analyzing graph Laplacians and operators that behave like Laplacians. Our point of view differs from that found in the existing literature; specifically, our point of departure is a stochastic process framework that utilizes the characterization of diffusion processes via drift and diffusion terms. This yields a general kernel-free framework for analyzing graph Laplacians with shrinking neighborhoods. We use it to extend the pointwise results of Hein et al. (2007) to cover non-smooth kernels and introduce location-dependent bandwidths. Applying these tools we are able to identify the asymptotic limit for a variety of graphs constructions including kNN, r-neighborhood, and "self-tuning" graphs. We are also able to provide an analysis for Locally Linear Embedding (Roweis & Saul, 2000).

A practical motivation for our interest in graph Laplacians based on kNN graphs is that these can be significantly sparser than those constructed using kernels, even if they have the same limit. Our framework allows us to establish this limiting equivalence. On the other hand, we can also exhibit cases in which kNN graphs converge to a different limit than graphs constructed from kernels, thereby explaining some cases where kNN graphs perform poorly. Moreover, our framework allows us to generate new algorithms: in particular, by using location-dependent bandwidths we obtain a class of operators that have nice spectral convergence properties that parallel those of the normalized Laplacian in von Luxburg et al. (2008), but which converge to a different class of limits.

2. The Framework

Our work exploits the connections among diffusion processes, elliptic operators (in particular the weighted Laplace-Beltrami operator), and stochastic differential equations (SDEs). This builds upon the diffusion process viewpoint in Nadler et al. (2006). Critically, we make the connection to the drift and diffusion terms of a diffusion process. This allows us to present a kernel-free framework for analysis of graph Laplacians as well as giving a better intuitive understanding of the limit diffusion process.

We first give a brief overview of these connections and present our general framework for the asymptotic analysis of graph Laplacians as well as providing some relevant background material. We then introduce our assumptions and derive our main results on the limit operator for a wide range of graph construction methods. We use these to calculate asymptotic limits for some specific graph constructions.

2.1. Relevant Differential Geometry

Assume $\mathcal{M}$ is a smooth m-dimensional manifold embedded in $\mathbb{R}^b$, the extrinsic coordinate system. To identify the asymptotic infinitesimal generator of a diffusion process on this manifold, we will derive the drift and diffusion terms in normal coordinates at each point. We refer the reader to Boothby (1986) for an exact definition of normal coordinates. For our purposes it suffices to note that the normal coordinates are coordinates in $\mathbb{R}^m$ that behave roughly as if a neighborhood of x was projected onto the tangent plane at x. To link the extrinsic coordinates of a point y in a neighborhood of x and normal coordinates s, we have the relation

$$y - x = H_x s + L_x(ss^T) + O(\|s^3\|), \qquad (1)$$

where H_x is a linear isomorphism between the normal coordinates in R^m and the m-dimensional tangent plane T_x at x. L_x is a linear operator describing the curvature of the manifold and takes $m \times m$ positive semidefinite matrices into the space orthogonal to the tangent plane, $T_x^\perp$. More advanced readers will note that this statement is Gauss' lemma and H_x and L_x are related to the first and second fundamental forms.

We are most interested in limits involving the operator defined below:

Definition 1 (Weighted Laplace-Beltrami operator). *Given a smooth manifold $\mathcal{M}$, the* weighted Laplace-Beltrami operator with respect to the density q *is the second-order differential operator $\Delta_q := \Delta_{\mathcal{M}} - \frac{\nabla q^T}{q}\nabla$ where $\Delta_{\mathcal{M}} := div \circ \nabla$ is the unweighted Laplace-Beltrami operator. For functions $f \in C^\infty(\mathcal{M})$ with support contained in the interior of the manifold, it induces a smoothness functional by the relationship*

$$\langle f, \Delta_q f\rangle_{L(q)} = \|\nabla f\|^2_{L_2(q)}. \qquad (2)$$

2.2. Equivalence of Limiting Characterizations

We now establish the connections among elliptic operators, diffusions, SDEs, and graph Laplacians. We first show that elliptic operators define diffusion processes and SDEs and vice versa. An elliptic operator $\mathcal{G}$ is a second-order differential operator of the form

$$\mathcal{G}f(x) = \sum_{ij} a_{ij}(x)\frac{\partial^2 f(x)}{\partial x_i \partial x_j} + \sum_i b_i(x)\frac{\partial f(x)}{\partial x_i} + c(x)f(x),$$

where the $m \times m$ coefficient matrix $(a_{ij}(x))$ is positive semidefinite for all x. If we use normal coordinates for a manifold, we see that the weighted Laplace-Beltrami operator Δ_q is a special case of an elliptic operator with $(a_{ij}(x)) = I$, the identity matrix, $b(x) = \frac{\nabla q(x)}{q(x)}$, and $c(x) = 0$. Diffusion processes are related via a result by Dynkin which states that, given a diffusion process, the generator of the process is an elliptic operator. The (infinitesimal) generator $\mathcal{G}$ of a diffusion process X_t is defined as

$$\mathcal{G}f(x) := \lim_{t \to 0} \frac{\mathbb{E}_x f(X_t) - f(x)}{t}$$

when the limit exists and convergence is uniform over x. Here $\mathbb{E}_x f(X_t) = \mathbb{E}(f(X_t)|X_0 = x)$. A converse relation holds as well. The Hille-Yosida theorem characterizes when a linear operator, such as an elliptic operator, is the generator of a stochastic process. We refer the reader to Kallenberg (2002) for details.

A time-homogeneous stochastic differential equation (SDE) defines a diffusion process as a solution (when one exists) to the equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,$$

where X_t is a diffusion process taking values in $\mathbb{R}^d$. The terms $\mu(x)$ and $\sigma(x)\sigma(x)^T$ are the *drift* and *diffusion* terms of the process.

By Dynkin's result, the generator $\mathcal{G}$ of this process is an elliptic operator and a simple calculation shows the operator is

$$\mathcal{G}f(x) = \frac{1}{2}\sum_{ij}\left(\sigma(x)\sigma(x)^T\right)_{ij}\frac{\partial^2 f(x)}{\partial x_i \partial x_j} + \sum_i \mu_i(x)\frac{\partial f(x)}{\partial x_i}.$$

All that remains then is to connect diffusion processes in continuous space to graph Laplacians on a finite set of points. Diffusion approximation theorems provide this connection. We state one version of such a theorem, which may be derived from Theorems 1.2.4, 1.6.3, and 7.4.2 in Ethier & Kurtz (1986).

Theorem 2 (Diffusion Approximation)**.** *Let $\mu(x)$ and $\sigma(x)\sigma(x)^T$ be drift and diffusion terms for a diffusion process defined on a compact set $S \subset \mathbb{R}^b$, and let G be the corresponding infinitesimal generator. Let $\{Y_t^{(n)}\}_t$ be Markov chains with transition matrices P_n on state spaces $\{x_i\}_{i=1}^n$ for all n, and let $c_n \uparrow \infty$ define a sequence of scalings. Put*

$$\begin{aligned}\hat{\mu}_n(x_i) &= c_n \mathbb{E}(Y_1^{(n)} - x_i | Y_0^{(n)} = x_i)\\ \hat{\sigma}_n(x_i)\hat{\sigma}_n(x_i)^T &= c_n \mathrm{Var}(Y_1^{(n)} | Y_0^{(n)} = x_i).\end{aligned}$$

Let $f \in C^3(S)$. If for all $\epsilon > 0$

$$\begin{aligned}\sup_{i \le n} \|\hat{\mu}_n(x_i) - \mu(x_i)\|_\infty &\to 0,\\ \sup_{i \le n} \left\|\hat{\sigma}_n(x_i)\hat{\sigma}_n(x_i)^T - \sigma(x_i)\sigma(x_i)^T\right\|_\infty &\to 0,\\ c_n \sup_{i \le n} \mathrm{P}\left(\left\|Y_1^{(n)} - x_i\right\| > \epsilon \middle| Y_0^{(n)} = x_i\right) &\to 0,\end{aligned}$$

then for the generators $A_n = c_n(P_n - I)$, we have $A_n f \to Gf$.

We remark that though the results we have discussed thus far are stated in the context of the extrinsic coordinates $\mathbb{R}^b$, we describe appropriate extensions in terms of normal coordinates in Ting et al. (2010).

2.3. Our Assumptions

We describe here the assumptions and notation for the rest of the paper. We will refer to the following assumptions as the *standard assumptions*.

Assume $\mathcal{M}$ is a smooth m-dimensional manifold isometrically embedded in $\mathbb{R}^b$. We further assume for simplicity that the manifold is compact without boundary, but describe weaker conditions in Ting et al. (2010). Unless stated explicitly otherwise, let f be an arbitrary function in $C^3(\mathcal{M})$.

Assume points $\{x_i\}_{i=1}^\infty$ are sampled i.i.d. from a density $p \in C^3(\mathcal{M})$ with respect to the natural volume element of the manifold, and assume that p is bounded away from zero.

For brevity, we will always use $x, y \in \mathbb{R}^b$ to be points on $\mathcal{M}$ expressed in extrinsic coordinates and let $s \in \mathbb{R}^m$ denote the normal coordinates for y in a neighborhood centered at x. Since they represent the same point, we will also use y and s interchangeably as function arguments; i.e., $f(y) = f(s)$. Whenever we take a gradient, it is with respect to the normal coordinates.

We define what we mean by convergence. When we write $g_n \to g$ where $domain(g_n) = \mathcal{X}_n \subset \mathcal{M}$, we mean $\|g_n - \pi_n g\|_\infty \to 0$ where $\pi_n g = g|_{\mathcal{X}_n}$ is the restriction of g to $\mathcal{X}_n$. For operators T_n on functions with domain $\mathcal{X}_n$, we take $T_n g = T_n \pi_n g$. Convergence of operators $T_n \to T$ means $T_n f \to Tf$ for all $f \in C^3(\mathcal{M})$.

We now introduce our assumptions on the graph construction methods. Let $K_0 : \mathbb{R}_+ \to \mathbb{R}_+$ be a base kernel with bounded variation and compact support. Let h_n be a sequence of bandwidth scalings and $r_x^{(n)}(\cdot) > 0, w_x^{(n)}(\cdot) \geq 0$ be (possibly random) location dependent bandwidth and weight functions that converge to $r_x(\cdot)$ and $w_x(\cdot)$, respectively, and have Taylor-like expansions for all $x, y \in \mathcal{M}$ with $\|x-y\| < h_n$:

$$r_x^{(n)}(y) = r_x(x) + (\dot{r}_x(x) + \alpha_x \text{sign}(u_x^T s) u_x)^T s + \epsilon_r^{(n)}(x,s)$$

$$w_x^{(n)}(y) = w_x(x) + \nabla w_x(x)^T s + \epsilon_w^{(n)}(x,s)$$

where the approximation error is uniformly bounded:

$$\sup_{x \in \mathcal{M}, \|s\| < h_n} |\epsilon_i^{(n)}(x,s)| = o(h_n^2) \quad \text{for } i = r, w.$$

We consider the limit of the random walk Laplacian defined by as $L_{rw} = I - D^{-1}W$ where I is the identity, W is the matrix of edge weights, and D is the diagonal degree matrix.

2.4. Main Theorem

Our main result is stated in the following theorem.

Theorem 3. *Assume the standard assumptions hold with probability 1. If the bandwidth scalings h_n satisfy $h_n \downarrow 0$, $nh_n^{d+2}/\log n \to \infty$, then for graphs constructed using the kernels*

$$K_n(x,y) = w_x^{(n)}(y) K_0\left(\frac{\|y-x\|}{h_n r_x^{(n)}(y)}\right) \tag{3}$$

there exists a constant $Z_{K_0,m}$ depending only on K_0 and the dimension m such that for $c_n = Z_{K_0,m}/h^{m+2}$,

$$-c_n L_{rw}^{(n)} f \to Af,$$

where A is the infinitesimal generator of a diffusion process with, in normal coordinates, drift and diffusion terms

$$\mu_s(x) = r_x(x)^2 \left(\frac{\nabla p(x)}{p(x)} + \frac{\nabla w(x)}{w(x)} + (m+2)\frac{\dot{r}_x(x)}{r_x(x)}\right),$$

$$\sigma_s(x)\sigma_s(x)^T = r_x(x)^2 I,$$

where I is the $m \times m$ identity matrix.

Note that the kernel in Eq. (3) generalizes previously analyzed graph constructions by (1) allowing for nonsmooth kernels, (2) introducing a random location-dependent bandwidth function $r_x(y)$, and (3) considering a general random weight function $w_x(y)$.

Proof. By the diffusion approximation theorem (Theorem 2) and since $h_n \downarrow 0$, we simply need to show uniform convergence of the drift and diffusion terms. We sketch the proof here and present the details in Ting et al. (2010). First, assuming we are given the true density p and limit weight and bandwidth functions, we calculate the limits assuming K_0 is an indicator kernel. To generalize to kernels of bounded variation and compact support, note that $K_0(x) = \int \mathbb{I}(|x| < z) d\eta_+(z) - \int \mathbb{I}(|x| < z) d\eta_-(z)$ for some finite positive measures η_-, η_+ with compact support. The result for general kernels then follows from Fubini's theorem.

The key calculation is establishing that integrating against an indicator kernel is like integrating over a sphere re-centered on $h_n^2 \dot{r}_x(x)$. Given this calculation and by Taylor expanding the non-kernel terms, one obtains the infinitesimal first and second moments and the degree operator:

$$\begin{aligned} M_1^{(n)}(x) &= \int s K_n(x,y) p(y) ds \\ &= C_{K_0,m} h_n^{m+2} r_x(x)^{m+2} \Bigg(w_x(x) \frac{\nabla p(x)}{m+2} + \\ &\quad + p(x) \frac{\nabla w_x(x)}{m+2} + w_x(x) p(x) \dot{r}_x(x) + O(h_n) \Bigg), \end{aligned}$$

$$M_2^{(n)}(x) = \frac{C_{K_0,m}}{m+2} h_n^{m+2} r_x(x)^{m+2} \left(w_x(x) p(x) I + O(h_n) \right),$$

$$d^{(n)}(x) = C'_{K_0,m} h_n^m r_x(x)^m \left(w(x) p(x) + O(h_n) \right),$$

for some constants $C_{K_0,m}, C'_{K_0,m}$.

Let $c_n = h_n^{-(m+2)} \frac{(m+2)C'_{K_0,m}}{C_{K_0,m}}$. Since K_n/d_n define Markov transition kernels, taking the limits $\mu_s(x) = \lim_{n\to\infty} c_n M_1^{(n)}(x)/d^{(n)}(x)$ and $\sigma_s(x)\sigma_s(x)^T = \lim_{n\to\infty} c_n M_2^{(n)}(x)/d^{(n)}(x)$ gives the stated result.

For the covergence of the empirical quantities, we find deterministic weight and bandwidth functions that upper and lower bound the moments. We may then apply Bernstein's inequality for i.i.d bounded random variables to obtain a.s. uniform convergence. □

2.5. Unnormalized and Normalized Laplacians

While our results are for the limit of the random walk Laplacian $L_{rw} = I - D^{-1}W$, it is easy to generalize them to the unnormalized Laplacian $L_u = D - W = DL_{rw}$ and symmetrically normalized Laplacian $L_{norm} = I - D^{-1/2}WD^{-1/2} = D^{1/2}L_{rw}D^{-1/2}$. For proof details see Ting et al. (2010).

Corollary 4. *Take the assumptions and definitions in Theorem 3, so that $c_n L_{rw}^{(n)} f \to Af$. Under the same assumptions on the bandwidth scaling, the rescaled degree terms $d^{(n)}(\cdot)/h_n^m$ converge uniformly a.s. to a* *function* $d(\cdot)$, *and*

$$c_n L_u^{(n)} f \to d \cdot Af \quad a.s.$$

Furthermore, if $d \in C^3(\mathcal{M})$ *and* $\frac{nh_n^{m+4}}{\log n} \to \infty$ *then*

$$c_n L_{norm}^{(n)} f \to d^{1/2} \cdot A(d^{-1/2} f) \quad a.s.$$

2.6. As weighted Laplace-Beltrami operator

Under some regularity conditions, the limit given in the main theorem (Theorem 3) yields a weighted Laplace-Beltrami operator. For convenience, define $\gamma(x) = r_x(x)$, $\omega(x) = w_x(x)$.

Corollary 5. *Assume the conditions of Theorem 3 and let* $q = p^2 \omega \gamma^{m+2}$. *If* $r_x(y) = r_y(x), w_x(y) = w_y(x)$ *for all* $x, y \in \mathcal{M}$ *and* $r_{(\cdot)}(\cdot), w_{(\cdot)}(\cdot)$ *are twice differentiable in a neighborhood of* (x, x) *for all* x, *then*

$$-c_n L_u^{(n)} \to \frac{q}{p} \Delta_q. \tag{4}$$

3. Application to Specific Graph Constructions

To illustrate Theorem 3, we apply it to calculate the asymptotic limits of graph Laplacians for several widely used graph construction methods. We also apply the general diffusion theory framework to analyze locally linear embedding.

3.1. r-Neighborhood and Kernel Graphs

In the case of the r-neighborhood graph, the base kernel is the indicator function $K_0(x) = I(|x| < r)$. The radius $r_x(y)$ is constant so $\dot{r}_x(x) = 0$. The drift is given by $\mu_s(x) = \nabla p(x)/p(x)$ and the diffusion term is $\sigma_s(x)\sigma_s(x)^T = I$. The limit operator is thus

$$\frac{1}{2}\Delta_{\mathcal{M}} + \frac{\nabla p(x)^T}{p(x)} \nabla = \frac{1}{2}\Delta_2$$

as expected. This analysis also holds for arbitrary kernels of bounded variation. One may also introduce the usual weight function $w_x^{(n)}(y) = d_n(x)^{-\alpha} d_n(y)^{-\alpha}$ to obtain limits of the form $\frac{1}{2}\Delta_{p^{(2-2\alpha)}}$. These limits match those obtained by Hein et al. (2007) and Lafon (2004) for smooth kernels.

3.2. Directed k-Nearest Neighbor Graph

kNN graphs are of particular interest since they produce very sparse graphs where the sparseness is easily controlled. For kNN graphs, the base kernel is still the indicator kernel, and the weight function is the constant 1. However, the bandwidth function $r_x^{(n)}(y)$ is random and depends on x. Since the graph is directed, it does not depend on y so $\dot{r}_x = 0$.

By the analysis in Section 3.4, $r_x(x) = cp^{-1/m}(x)$ for some constant c. Consequently the limit operator is proportional to

$$\frac{1}{p^{2/m}}(x)\left(\Delta_{\mathcal{M}} + 2\frac{\nabla p^T}{p}\nabla\right) = \frac{1}{p^{2/m}}\Delta_{p^2}.$$

Note that this is *not* a self-adjoint operator in $L(p)$. The symmetrization of the graph has a non-trivial effect to make the graph Laplacian self-adjoint.

3.3. Undirected k-Nearest Neighbor Graph

We consider the "OR-construction" where nodes v_i and v_j are linked if v_i is a k^{th}-nearest neighbor of v_j *or* vice-versa. In this case $h_n^m r_x^{(n)}(y) = \max\{\rho_n(x), \rho_n(y)\}$ where $\rho_n(x)$ is the distance to the k_n^{th} nearest neighbor of x. The limit bandwidth function is non-differentiable, $r_x(y) = \max\{p^{-1/m}(x), p^{-1/m}(y)\}$, but a Taylor-like expansion exists with $\dot{r}_x(x) = -\frac{1}{2m}\frac{\nabla p(x)^T}{p(x)}$. The limit operator is

$$\frac{1}{p^{2/m}}\Delta_{p^{1-2/m}},$$

which is self-adjoint in $L_2(p)$. Surprisingly, if $m = 1$ then the kNN graph construction induces a drift *away* from high density regions.

3.4. Conditions for kNN convergence

To complete the analysis, we must check that the conditions for kNN graph constructions satisfy the assumptions of the main theorem. This is a straightforward application of existing uniform consistency results for kNN density estimation.

Let $h_n = \left(\frac{k_n}{n}\right)^{1/m}$. The condition we must verify is

$$\sup_{y \in \mathcal{M}} \left\| r_x^{(n)} - r_x \right\|_\infty = O(h_n^2) \text{ a.s.}$$

We check this for the directed kNN graph, but analyses for other kNN graphs are similar. The kNN density estimate of Loftsgaarden & Quesenberry (1965) is

$$\hat{p}_n(x) = \frac{k_n}{nV_m(h_n r_x^{(n)}(x))^m}, \tag{5}$$

where $h_n r_x^{(n)}(x)$ is the distance to the k_n^{th} nearest neighbor of x given n data points and V_m is the volume of the m-dimensional unit sphere. Taylor-expanding Eq. (5) shows that if $\|\hat{p}_n - p\|_\infty = O(h_n^2)$ a.s. then the requirement on the location-dependent bandwidth for the main theorem is satisfied.

Devroye & Wagner (1977)'s proof for the uniform consistency of kNN density estimation may be easily modified to show this. Take $\epsilon = (k_n/n)^2$ in their proof. One then sees that $h_n = k_n/n \to 0$ and $\frac{nh_n^{m+2}}{\log n} = \frac{k_n^{2+2/m}}{n^{1+2/m}\log n} \to \infty$ are sufficient to achieve the desired bound on the error.

3.5. "Self-Tuning" Graphs

The form of the kernel used in self-tuning graphs is

$$K_n(x,y) = \exp\left(\frac{-\|x-y\|^2}{\sigma_n(x)\sigma_n(y)}\right),$$

where $\sigma_n(x) = \rho_n(x)$, the distance between x and the k^{th} nearest neighbor. The limit bandwidth function is $r_x(y) = \sqrt{p^{-1/m}(x)p^{-1/m}(y)}$. Since this is twice differentiable, Corollary 5 gives the asymptotic limit, which is the same as for undirected kNN graphs:

$$p^{-2/m}\Delta_{p^{1-2/m}}.$$

3.6. Locally Linear Embedding

Locally linear embedding (LLE), introduced by Roweis & Saul (2000), has been noted to behave like (the square of) the Laplace-Beltrami operator (Belkin & Niyogi, 2003).

Using our kernel-free framework we will show how LLE differs from weighted Laplace-Beltrami operators and graph Laplacians in several ways.

The key observation is that LLE only controls for the drift term in the extrinsic coordinates. Thus, the diffusion term has freedom to vary. However, if the manifold has curvature, the drift in extrinsic coordinates constrains the diffusion term in normal coordinates.

The LLE matrix is defined as $(I-W)^T(I-W)$ where W is a weight matrix which minimizes reconstruction error $W = \text{argmin}_{W'} \|(I-W')y\|^2$ under the constraints that $W'1 = 1$ and $W'_{ij} \neq 0$ only if j is one of the k^{th} nearest neighbors of i. Typically $k > m$ and reconstruction error $= 0$. We will assume this and analyze the matrix $M = W - I$.

Suppose LLE produces a sequence of matrices $M_n = I - W_n$. The row sums of each matrix are zero. Thus, we may decompose $M_n = A_n^+ - A_n^-$ where A_n^+, A_n^- are infinitesimal generators for finite state Markov processes obtained from the positive and negative weights respectively. Assume that there is some scaling c_n such that $c_nA_n^+, c_nA_n^-$ converge to generators of diffusion processes with drifts μ_+, μ_- and diffusion terms $\sigma_+\sigma_+^T, \sigma_-\sigma_-^T$. Set $\mu = \mu_+ - \mu_-$ and $\sigma\sigma^T = \sigma_+\sigma_+ - \sigma_-\sigma_-$.

Since reconstruction error in extrinsic coordinates is zero, Eq. (1) shows that in normal coordinates we have

$$\mu_s(x) = 0 \quad \text{and} \quad L_x(\sigma_s(x)\sigma_s(x)^T) = 0.$$

From this we see that: (1) LLE can only behave like an *unweighted* Laplace-Beltrami operator since the drift term is zero. (2) LLE is affected by the curvature of the manifold since $\sigma_s(x)\sigma_s(x)^T$ must lie in the null space of L_x. Furthermore, when L_x is full rank then $\sigma_s = 0$ and LLE may behave unlike any elliptic operator (including the Laplace-Beltrami operator). (3) LLE has, in general, *no well-defined asymptotic limit* without additional conditions on the weights, since the diffusion term is free to vary in the null space of L_x.

We note that while the LLE framework of minimizing reconstruction error can yield ill-behaved solutions, practical implementations add a regularization term when constructing the weights. This causes the reconstruction error to be non-zero in general and gives unique solutions for the weights that favor equal weights (and hence asymptotic behavior akin to that of kNN graphs).

4. Experiments

To illustrate the theory, we show how to correct the bad behavior of the kNN Laplacian for a synthetic data set. We also show how our analysis can predict the surprising behavior of LLE.

kNN Laplacian. We first consider an example which almost all non-linear embedding techniques handle well but where the kNN graph Laplacian performs poorly. Figure 1 shows a 2D manifold embedded in three dimensions and shows embeddings using different graph constructions. The theoretical limit of the normalized Laplacian L_{knn} for a kNN graph is $L_{knn} = \frac{1}{p}\Delta_1$, while the limit for a graph with Gaussian weights is $L_{gauss} = \Delta_p$. The first two coordinates of each point are from a truncated normal distribution, so the density at the boundary is small and the effect of the $1/p$ term is substantial. This yields the bad behavior shown in Figure 1 (C). We may use Eq. (5) as a pilot estimate $\hat{p}$ of the density. Choosing $w_x(y) = \sqrt{\hat{p}_n(x)\hat{p}_n(y)}$ gives a weighted kNN graph with the same limit as the graph with Gaussian weights. Figure 1(D) shows that this change yields the roughly desired behavior but with fewer "holes" in low density regions and more "holes" in high density regions.

LLE. We consider another synthetic data set, the toroidal helix, in which the manifold structure is easy to recover. Figure 2(A) shows the manifold which is clearly isomorphic to a circle, a fact picked up by the

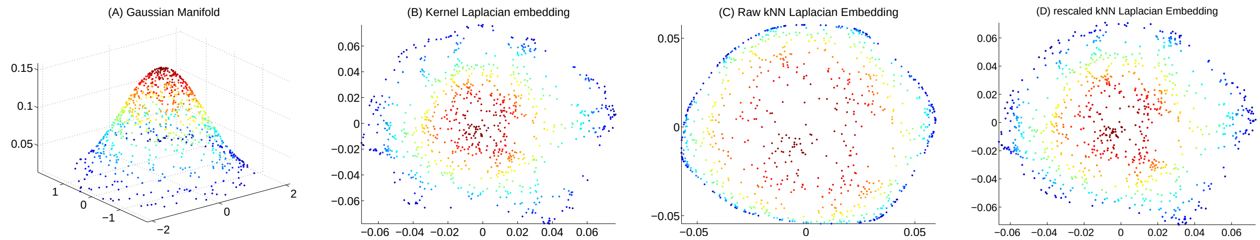

Figure 1. (A) shows a 2D manifold where the x and y coordinates are drawn from a truncated standard normal distribution. (B-D) show embeddings using different graph constructions. (B) uses a normalized Gaussian kernel $\frac{K(x,y)}{d(x)^{1/2}d(y)^{1/2}}$, (C) uses a kNN graph, and (D) uses a kNN graph with edge weights $\sqrt{\hat{p}(x)\hat{p}(y)}$. The bandwidth for (B) was chosen to be the median standard deviation from taking 1 step in the kNN graph in order to have comparable sparsity.

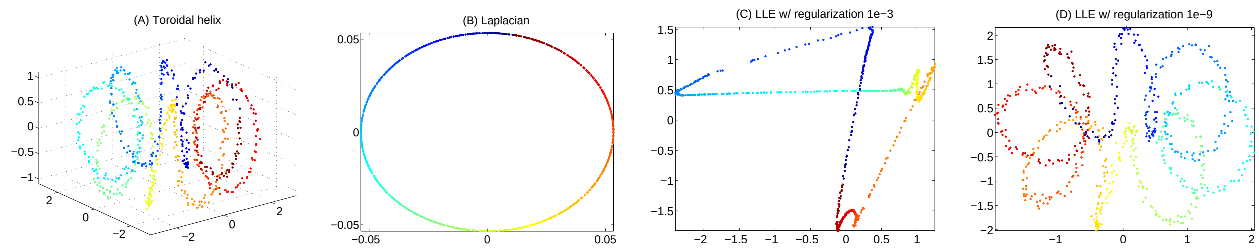

Figure 2. (A) shows a 1D manifold isomorphic to a circle. (B-D) show the embeddings using (B) Laplacian eigenmaps which correctly identifies the structure, (C) LLE with regularization 1e-3, and (D) LLE with regularization 1e-6.

kNN Laplacian in Figure 2(B).

Our theory predicts that the heuristic argument that LLE behaves like the Laplace-Beltrami operator will *not* hold. Since the total dimension for the drift and diffusion terms is two and the extrinsic coordinates also have dimension three, that there is forced cancellation of the first- and second-order differential terms and the operator should behave like the zero operator or include higher-order differentials. In Figure 2(C) and (D), we see this that LLE performs poorly and that the behavior comes closer to the zero operator when the regularization term is smaller.

5. Remarks

Convergence rates. We note that one missing element in our analysis is the derivation of convergence rates. For the main theorem, we note that although we have employed a diffusion approximation theorem it is not in fact necessary to use such a theorem. Since our theorem still uses a kernel (albeit one with much weaker conditions), a virtually identical proof can be obtained by applying a function f and Taylor-expanding it. Thus, we believe that similar convergence rates to those in Hein et al. (2007) can be obtained. Also, while our convergence result is stated for the strong operator topology, the same conditions as in Hein should give weak operator convergence.

Eigenvalues/eigenvectors. Eigenvectors of the Laplacian matrix are of particular interest since they may be used as a set of harmonic basis functions or in spectral clustering.

With our more general graph constructions we can use the theory of compact integral operators to obtain graph Laplacians that (1) have eigenvectors that converge for fixed (non-decreasing) bandwidth scalings and (2) converge to a limit that is different from that of previously analyzed normalized Laplacians when the bandwidth decreases to zero. In particular, for arbitrary $q, g \in C^3(\mathcal{M})$ with g bounded away from zero. We may choose w, r such that, if $h_n \downarrow 0$ at an appropriate rate, we obtain limits of the form

$$-c_n L^{(n)}_{norm} f \to g^{-1/2}\frac{q}{p}\Delta_q(g^{-1/2}f),$$

with corresponding smoothness functional

$$\left\langle f, g^{-1/2}\frac{q}{p}\Delta_q(g^{-1/2}f)\right\rangle_{L_2(p)} = \left\|\nabla(g^{-1/2}f)\right\|^2_{L_2(q)},$$

and if $h_n = h_1$ is constant then the eigenvectors of $L^{(n)}_{norm}$ converge in the sense given by von Luxburg et al. (2008).

6. Conclusions

We have introduced a general framework that enables us to analyze a wide class of graph Laplacian constructions. Our framework reduces the problem of graph Laplacian analysis to the calculation of a mean and variance (or drift and diffusion) for any graph construction method with positive weights and shrinking neighborhoods. Our main theorem extends existing strong operator convergence results to non-smooth kernels, and introduces a general location-dependent bandwidth function. The analysis of a location-dependent bandwidth function, in particular, significantly extends the family of graph constructions for which an asymptotic limit is known. This family includes the previously unstudied (but commonly used) kNN graph constructions, unweighted r-neighborhood graphs, and "self-tuning" graphs.

Our results also have practical significance in graph constructions as they suggest graph constructions that (1) produce sparser graphs than those constructed with the usual kernel methods, despite having the same asymptotic limit, and (2) in the fixed bandwidth regime, produce unnormalized Laplacians or normalized Laplacians that have well-behaved spectra but converge to a different class of limit operators than previously studied normalized and unnormalized Laplacians. In particular, this class of limits include those that induce the smoothness functional $\|\nabla f\|^2_{L_2(q)}$ for almost any density q. The graph constructions may also (3) have better connectivity properties in low-density regions.

Acknowledgements. We would like to thank Martin Wainwright and Bin Yu for their helpful comments, and our anonymous reviewers for the detailed and helpful review.

References

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Belkin, M. and Niyogi, P. Towards a theoretical foundation for Laplacian-based manifold methods. In *COLT*. Springer, 2005.

Belkin, M. and Niyogi, P. Convergence of Laplacian eigenmaps. In *NIPS 19*, 2006.

Boothby, W. M. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 1986.

Bousquet, O., Chapelle, O., and Hein, M. Measure based regularization. In *NIPS 16*, 2003.

Devroye, L.P. and Wagner, T.J. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, pp. 536–540, 1977.

Ethier, S. and Kurtz, T. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

Giné, E. and Koltchinskii, V. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *4th International Conference on High Dimensional Probability*, 2005.

Hein, M., Audibert, J., and Von Luxburg, U. From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. In *COLT*, 2005.

Hein, M., Audibert, J.-Y., and von Luxburg, U. Graph Laplacians and their convergence on random neighborhood graphs. *JMLR*, 8:1325–1370, 2007.

Kallenberg, O. *Foundations of Modern Probability*. Springer Verlag, 2002.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

Lafon, S. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, CT, 2004.

Loftsgaarden, D.O. and Quesenberry, C.P. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Maier, M., von Luxburg, U., and Hein, M. Influence of graph construction on graph-based clustering measures. In *NIPS 21*, 2008.

Nadler, B., Lafon, S., Coifman, R., and Kevrekidis, I. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. In *Applied and Computational Harmonic Analysis*, 2006.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323, 2000.

Singer, A. From graph to manifold Laplacian: the convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

Ting, D., Huang, L., and Jordan, M. An analysis of the convergence of graph Laplacians. Technical report, University of California, Berkeley, 2010.

von Luxburg, U., Belkin, M., and Bousquet, O. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. In *NIPS 17*, 2004.

Zhu, Xiaojin. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin Madison, 2008.