Variable Bandwidth Diffusion Kernels

Tyrus Berry\textsuperscript{a,*}, John Harlim\textsuperscript{a,b}

\textsuperscript{a}Department of Mathematics, the Pennsylvania State University, 109 McAllister Building, University Park, PA 16802-6400, USA
\textsuperscript{b}Department of Meteorology, the Pennsylvania State University, 503 Walker Building, University Park, PA 16802-5013, USA

Abstract

A practical limitation of operator estimation via kernels is the assumption of a compact manifold. In practice we are often interested in data sets whose sampling density may be arbitrarily small, which implies that the data lies on an open set and cannot be modeled as a compact manifold. In this paper, we show that this limitation can be overcome by varying the bandwidth of the kernel spatially. We present an asymptotic expansion of these variable bandwidth kernels for arbitrary bandwidth functions; generalizing the theory of Diffusion Maps and Laplacian Eigenmaps. Subsequently, we present error estimates for the corresponding discrete operators, which reveal how the small sampling density leads to large errors; particularly for fixed bandwidth kernels. By choosing a bandwidth function inversely proportional to the sampling density (which can be estimated from data) we are able to control these error estimates uniformly over a non-compact manifold, assuming only fast decay of the density at infinity in the ambient space. We numerically verify these results by constructing the generator of the Ornstein-Uhlenbeck process on the real line using data sampled independently from the invariant measure. In this example, we find that the fixed bandwidth kernels yield reasonable approximations for small data sets when the bandwidth is carefully tuned, however these approximations actually degrade as the amount of data is increased. On the other hand, an operator approximation based on a variable bandwidth kernel does converge in the limit of large data; and for small data sets exhibits reduced sensitivity to bandwidth selection. Moreover, even for compact manifolds, variable bandwidth kernels give better approximations with reduced dependence on bandwidth selection. These results extend the classical statistical theory of variable bandwidth density estimation to operator approximation.

Keywords: diffusion maps, variable bandwidth kernels, manifold learning, nonparametric modeling

1. Introduction

Graph Laplacian and kernel based techniques are ubiquitous in machine learning, clustering, classification. While these practical algorithms have been very successful in various applications, they were not mathematically understood until the development of Laplacian Eigenmaps [1] and Diffusion Maps [4] as well as other works on the convergence of graph Laplacians to their continuous counterparts [10, 17]. The novel perspective taken by these authors was to construct a stochastic matrix whose generator, $L_{\epsilon,\alpha}$, is a discrete representation of a continuous, Kolmogorov operator,

\[
L_{\epsilon,\alpha} f = \Delta f + (2 - 2\alpha) \nabla f \cdot \nabla q / q,
\]

for arbitrarily smooth function $f$. The stochastic matrix is constructed by evaluating a homogeneous kernel,

\[
K_\epsilon(x, y) = h \left( \frac{||x - y||^2}{\epsilon} \right),
\]

with exponential decays in the distance $||x - y||$, on all pairs of points $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$, sampled from a $d$-dimensional manifold $M \subset \mathbb{R}^n$ with smooth sampling density $q$. Note that $\Delta$ is the Laplacian and $\nabla$ is the gradient operator,

*Corresponding author

Email addresses: tvb11@psu.edu (Tyrus Berry), jharlim@psu.edu (John Harlim)
and both operators are defined with respect to the Riemannian metric that the manifold $\mathcal{M}$ inherits from the ambient space $\mathbb{R}^n$. We should clarify that in this paper, we refer to the Laplacian operator as the negative of Laplace-Beltrami operator for convenience. The parameter $\alpha$ controls the degree for which the sampling distribution is allowed to bias the operator, and a key result of [4] is that setting $\alpha = 1$ removes the bias entirely and recovers the Laplacian operator independent of the sampling density $q$. Moreover, setting $\alpha = 1/2$ recovers the backward Kolmogorov operator of a gradient flow with potential $U = -\log q$, and setting $\alpha = 0$ recovers the normalized graph Laplacian on a graph with isotropic (Gaussian) weights [1] which approximates the Laplacian operator when the sampling density is uniform. For the case of $\alpha = 0$ with uniform sampling, the error estimates of Singer [15] showed that for any sufficiently smooth function $f$ at any point $x_i$ in the data set,

$$L_{\epsilon,\alpha} f(x_i) \equiv \frac{1}{\epsilon} \left( \frac{\sum_j K(x_i, x_j) f(x_j)}{\sum_j K(x_i, x_j)} - f(x_i) \right) = L_{\alpha} f(x_i) + O \left( \epsilon, \frac{||\nabla f(x)||}{\sqrt{N\epsilon^{1/2}d/4}} \right)$$

with high probability. The discrete operator $L_{\epsilon,\alpha}$ is closely related to the graph Laplacian with edge weights $w_{ij} = K(x_i, x_j)$ given by the kernel $K$ (see [1, 4]).

These discrete approximations of operators have two important applications based on using the eigenvectors of the discrete operator as approximations to the eigenfunctions of the continuous operator. First, as shown by Coifman and Lafon in [4], the eigenfunctions form the components of a nonlinear map into a low-dimensional Euclidean space called a diffusion map. Diffusion maps gave a new mathematical interpretation for many algorithms used in machine learning as well as clustering algorithms based on kernel principal components or weighted graph Laplacians. Second, when the data is generated by a dynamical system, the eigenfunctions of the continuous operator $L$ can be used to find low-dimensional representations of the long-time dynamics as shown in [6, 3, 5]. These ideas were generalized to time-delay reconstructions of dynamical systems in the work of [2] and related work in [7, 8].

One practical limitation of the kernel based operator estimation theory is the fundamental assumption that the data must lie on a compact manifold. For a non-compact manifold where the sampling is not bounded away from zero, we will see that a fixed bandwidth kernel can only give reasonable approximations for small data sets and a carefully tuned bandwidth. Indeed, the discrete operator approximations based on fixed bandwidth kernels do not converge to the continuous operator in the limit of large data. In this paper, we will show that this limitation can be overcome by extending the theory of Diffusion Maps to non-homogeneous kernels of the form,

$$K^\alpha(x, y) = h \left( \frac{||x - y||^2}{\epsilon \rho(x) \rho(y)} \right),$$

(4)

with a bandwidth function, $\rho$. This choice was motivated by the problem of bandwidth selection, which is a well known issue in kernel density estimation problems [13, 12] for which the variable bandwidth kernels were advocated for estimating tails of distribution with sparse sampling data set, see for example [16, 14]. Independently, kernels of the form (4) have been proposed in [7, 8] for describing data generated by a dynamical system. They chose a bandwidth function, $\rho$, based on the distance traveled in state space on a fixed time unit. Combined with time-delay embeddings, this bandwidth function was shown to give a natural scaling which reduces the dependence on the initial observation function. Although their intuition is appropriate for describing chaotic dynamical systems with intermittent regime transitions, the technique of [7, 8] is constrained to dynamical systems on compact manifolds.

In this paper, we develop a general theory for arbitrary bandwidth functions in the form of (4) and focus on a different choice of bandwidth function than in [7, 8]. Using such variable bandwidth kernels, our goal is to approximate the weighted diffusion operator of data sets sampled from open sets, assuming that the sampling density, $q$, rapidly decays at infinity. Following [4, 1], the intrinsic differential operator will be estimated without any prior knowledge of the manifold structure or data distribution. In fact the continuous theory developed in [4] trivially extends to these non-compact manifolds, however by extending the error estimate in (3) of [15], we will find the errors in the discrete approximation of the continuous operators are unbounded in areas of sparse sampling. By choosing the bandwidth function $\rho$ to be inversely proportional to the sampling density, we are able to control the error bounds in the areas of sparse sampling. Intuitively, this corresponds to a variable bandwidth kernel where the bandwidth is large in areas of sparse sampling and the bandwidth is small in areas of dense sampling.

Our main contribution in this paper can be summarized as follows:
Theorem 1. Let $q \in L^1(\mathcal{M}) \cap C^3(\mathcal{M})$ be a density that is bounded above on an embedded $d$-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ and let $(x_i)_{i=1}^N$ be sampled independently with distribution $q$. Let $K^q_{\rho}(x,y)$ be a variable bandwidth kernel of the form (4) with bandwidth function $\rho = q^2 + O(\epsilon)$ and shape function $h : [0, \infty) \to [0, \infty)$ with exponential decay at infinity. For a smooth real-valued function $f \in L^2(\mathcal{M}, q) \cap C^3(\mathcal{M})$ and an arbitrary point $x_i \in \mathcal{M}$, define the discrete functionals,

$$F_i(x_j) = \frac{K^q_{\rho}(x_i, x_j)f(x_j)}{q^2_{\rho}(x_j)q^2_{\rho}(x_i)^{d/2}}, \quad G_i(x_j) = \frac{K^q_{\rho}(x_i, x_j)}{q^2_{\rho}(x_j)q^2_{\rho}(x_i)^{d/2}},$$

where $q^2_{\rho}(x_i) = \sum_k K^q_{\rho}(x_i, x_k)/q(x_k)^{d/2}$ is a kernel density estimate of the sampling density $q$. Then, with high probability,

$$L_{\epsilon, \alpha, \beta} f(x_i) = \frac{1}{\text{emp}(x_i)^2} \left( \sum_j F_i(x_j) - f(x_i) \right) = \mathcal{L}_{\alpha, \beta} f(x_i) + O \left( \epsilon, q(x_i)^{(1-d)/2} \right),$$

for some finite valued constant $m$, where

$$\mathcal{L}_{\alpha, \beta} f \equiv \Delta f + c_1 \nabla f \cdot \frac{\nabla q}{q}, \quad (5)$$

$c_1 = 2 - 2\alpha + d\beta + 2\beta$ and $c_2 = 1/2 - 2\alpha + 2d\alpha + d\beta/2 + \beta$.

Notice that the bandwidth function above is $\rho = q^2 + O(\epsilon)$ which means we do not require that the sampling density is exactly known. For practical applications we may use any kernel density estimation to find an order-$\epsilon$ approximation of $q$ for the purposes of defining the bandwidth function $\rho$. However, we note that the normalization term $q^2_{\rho}$ may not be replaced with an alternate density estimate because the result in Theorem 1 carefully accounts for the higher order terms in the asymptotic expansion of $q^2_{\rho}$.

The error bound in Theorem 1 has three components. The $O(\epsilon)$ component is due to the error between the true operator $\mathcal{L}_{\alpha, \beta}$ and the operator $L_{\epsilon, \alpha, \beta}$ with the summations replaced by the expectations, where $\mathbb{E}[f] \equiv \int_{\mathcal{M}} f(z)q(z)d\mathcal{V}(z)$. In other words, for fixed $\epsilon$, in the limit of large $N$ the error will be of order $\epsilon$ assuming the first term dominates.

The error term $O \left( \epsilon, q(x_i)^{(1-d)/2} \right)$ is due to the need to obtain an order-$\epsilon^2$ estimate of $q^2_{\rho}$. While this term dominates for $q = O(1)$, as $q \to 0$ the third component of the error may become dominant. The error term $O \left( \frac{(||\nabla f||_x||q||^2_{x_i})^{d/2}}{\sqrt{N^{d-1/2}\epsilon}} \right)$ is due to the bias error between the ratio of discrete sums $\sum_j F_i(x_j)/\left( \sum_j G_i(x_j) \right)$ and the continuous expectations $\mathbb{E}[F_i]/\mathbb{E}[G_i]$. Depending on the choices of $\alpha$ and $\beta$, this term can dominate the error in areas of sparse sampling where $q$ is small. In particular, if $q$ is not bounded away from zero, as $N$ increases the data will begin to sample areas of small density, and this final error term can actually increase as the amount of data increases when $c_2 > 0$.

The $\alpha$ normalization in the functionals $F_i, G_j$ is a de-biasing parameter, which is equivalent to the diffusion maps $\alpha$ normalization so that when $\beta = 0$ we recover the operator $\mathcal{L}_{0, \alpha} = \mathcal{L}_\alpha$ in (1) since $c_1 = 2 - 2\alpha$. Notice that $\beta = 0$ is the one case which does not require knowledge of the intrinsic dimension $d$ of the manifold $\mathcal{M}$. However, when $\beta = 0, \alpha > 0$, and $d \in \mathbb{N}$, we have $c_2 = 1/2 + 2\alpha(d - 1) > 0$ which means that the error may be unbounded as $q \to 0$. This crucial observation explains why the the fixed bandwidth kernel in (2) is impractical when the sampling measure is not bounded away from zero. By taking $\beta < 0$ we can make $c_2 \leq 0$ which implies the point-wise errors are uniformly bounded and we will recover the continuous operator in the limit of large data. Taking $\beta < 0$ will require knowledge of the intrinsic dimension $d$ of $\mathcal{M}$, see [9, 11] for some methods and considerations for estimating the intrinsic dimension.

A related issue on non-compact manifolds is that $||\nabla f||$ may be unbounded. In particular, in Section 5 we consider the Hermite polynomials which are eigenfunctions of the Kolmogorov operator of a stochastically forced gradient flow with a quadratic potential on the real line. As long as $q$ has sufficiently fast decay at infinity and $c_2 < 0$, the term $q^{-c_2}$ will control the growth of $||\nabla f||$ to allow for uniform point-wise error bounds on the data set. It would require an infinite amount of data to construct the entire operator on an unbounded domain, however, for a finite amount of data we can correctly estimate the operator point-wise with bounded error over the entire data set by taking $\beta < 0$ to sufficiently force $c_2 < 0$.

The remainder of this paper is organized as follows: In Section 2 we present the continuous theory of variable bandwidth kernels for operator estimation, generalizing the theory of [4] to kernels of the form (4). In Section 3, we
generalize the error estimates of [15] to variable bandwidth kernels with non-uniform sampling, this will complete the proof of Theorem 1. In Section 4 we give the details of the numerical algorithm including some important numerical considerations for optimal implementation. In Section 5 we present a numerical example on an unbounded manifold which demonstrates and validates the results of Theorem 1. In Section 6 we present a numerical example on a compact domain which compares fixed and variable bandwidth kernels for estimating the eigenfunctions of the Laplacian operator. We conclude with a short summary in Section 7.

2. Variable Bandwidth Kernels

The goal of this section is to determine the rate of convergence of the generator of the integral operator, \( G_\epsilon^s = \epsilon^{-d/2} \int_M K_\epsilon^s(x, y)f(y) d\nu(y) \), associated with the kernel \( K_\epsilon^s \) in (4) to \( L_{a, b} \) in (5) as \( \epsilon \to 0 \). To achieve this goal, we need to determine the asymptotic expansion of \( G_\epsilon^s \) with respect to \( \epsilon \).

In order to find this expansion, we first extend a key technical lemma (cf. Lemma 2.1) of [4] to non-compact manifolds in Section 2.1. In Section 2.2, we find the expansion for the left formulation, with kernel,

\[
K_\epsilon^l(x, y) \equiv h \left( \frac{||x - y||^2}{\epsilon \rho(x)} \right),
\]

by applying a change of variables and then using the result in Section 2.1. In Section 2.3, using the expansion of the left formulation we use the weak formulation of an operator to find the expansion of the right formulation, with kernel,

\[
K_\epsilon^r(x, y) \equiv h \left( \frac{||x - y||^2}{\epsilon \rho(y)} \right).
\]

In Section 2.4, we combine these two results to find the expansion of the symmetric formulation. For the sake of clarity, we will compute all these expansions assuming uniform sampling and then we will extend the final expansion of \( G_\epsilon^s \) to non-uniform sampling in Section 2.5. Of course, the left and right formulations can also be extended to non-uniform sampling using the same technique.

2.1. Extending the asymptotic expansion of diffusion maps

In this section, we extend the following fundamental lemma for kernel operator estimation to non-compact manifolds.

**Lemma 2.1 (Expansion of Fixed Bandwidth Kernels, Coifman and Lafon [4]).** Let \( f \) be a smooth real-valued function on an embedded \( d \)-dimensional compact manifold \( M \subset \mathbb{R}^n \) and let \( h : [0, \infty) \to [0, \infty) \) have fast decay, then we have

\[
G_\epsilon f(x) = \epsilon^{-d/2} \int_M h \left( \frac{||x - y||^2}{\epsilon} \right) f(y) d\nu(y) = m_0 f(x) + \epsilon m_2 (\omega(x) f(x) + \Delta f(x)) + O(\epsilon^2)
\]

where \( m_0 = \int_{\mathbb{R}^d} h(||z||^2) dz \) and \( m_2 = \frac{1}{2} \int_{\mathbb{R}^d} \frac{1}{||z||^2} h(||z||^2) dz \) are constants determined by \( h \), and \( \omega \) depends on the induced geometry of \( M \). Note that \( d\nu(y) \) is the volume form on \( M \) and the operator \( \Delta \) is the (negative definite) Laplacian on \( M \), and these are both defined with respect to the Riemannian metric inherited from the ambient space.

The above lemma assumes uniform sampling on a compact manifold since we can estimate the operator \( G_\epsilon f \) by the discrete sum,

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} K_\epsilon(x, x_i)f(x_i) = \int_M K_\epsilon(x, y)f(y) q(y) d\nu(y) = \frac{1}{\text{vol}(M)} \int_M K_\epsilon(x, y)f(y) d\nu(y) = \frac{\epsilon^{d/2}}{\text{vol}(M)} G_\epsilon f(x),
\]

where the second equality follows from assuming \( q \) is sampled uniformly with respect to the volume form \( d\nu(y) \) on the manifold \( M \). To see this, let \( q(y) = c \) be uniform with respect to \( d\nu(y) \). Since \( q \) is a density, we have \( 1 = \int_M q(y) d\nu(y) = \int_M c d\nu(y) = c \text{vol}(M) \) and so \( q(y) = c = \text{vol}(M)^{-1} \) which explains why the Monte-Carlo sum estimates \( \frac{\epsilon^{d/2}}{\text{vol}(M)} G_\epsilon f \) instead of \( G_\epsilon f \).
where \( M \) is not compact the integral in \( G_\epsilon \) may diverge, however, for practical applications which sample a non-compact manifold, the sampling density is typically not uniform, and especially on unbounded manifolds we will assume fast decay of the sampling density at infinity. Here, we generalize the lemma of \([4]\) to non-compact manifolds as follows.

**Lemma 2.2** (Expansion of Fixed Bandwidth Kernels on non-Compact Manifolds). \( M \subset \mathbb{R}^n \) be an embedded \( d \)-dimensional manifold and let \( q : M \rightarrow (0, \infty) \) be bounded above such that \( q \in L^1(M) \cap C^3(M) \). Let \( h : [0, \infty) \rightarrow [0, \infty) \) have fast decay in the sense that there exists \( a, \sigma \) such that \( h(x) < a \exp(-x/\sigma) \). Then for all \( f \in L^2(M, q) \cap C^3(M) \) we have,

\[
G_\epsilon(f q)(x) = \epsilon^{-d/2} \int_M h \left( \frac{||x - y||^2}{\epsilon} \right) f(y) q(y) \, dV(y) = m_0 f(x) q(x) + \epsilon m_2 (\omega(x) f(x) q(x) + \Delta f q(x)) + O(\epsilon^2)
\]

Proof. We note that in the proof of Lemma 2.1 in \([4]\) compactness is only used to bound the integral outside of a neighborhood of radius \( \epsilon^2 \) for \( 0 < \gamma < 1/2 \) around \( x \). Since \( f \in L^2(M, q) \) we have,

\[
\epsilon^{-d/2} \int_{y \in M, ||x - y|| > \epsilon^2} h \left( \frac{||x - y||^2}{\epsilon} \right) f(y) q(y) \, dV(y) \leq ||f||_{L^2(M)} \epsilon^{-d/2} \int_{y \in M, ||x - y|| > \epsilon^2} h \left( \frac{||x - y||^2}{\epsilon} \right) q(y) \, dV(y)
\]

\[
= ||f||_{L^2(M)} \epsilon^{-d/2} \int_{y \in \hat{M}, \hat{||x - y||} > \gamma \epsilon^2} h(\gamma^2 \hat{||x - y||}^2) e^{\gamma^2 \hat{||x - \epsilon^{-1/2} z||} / \sigma} \, d\hat{V}(\hat{z})
\]

\[
\leq ||f||_{L^2(M)} ||q||_{L^\infty} \int_{\hat{z} \in \hat{M}, \hat{||x - \epsilon^{-1/2} z||} > \gamma \epsilon^2} a e^{-\gamma^2 \hat{||x - \epsilon^{-1/2} z||}^2 / \sigma} \, d\hat{V}(\hat{z})
\]

\[
\leq ||f||_{L^2(M)} ||q||_{L^\infty} \int_{\hat{z} \in \hat{M}, \hat{||x - \epsilon^{-1/2} z||} > \gamma \epsilon^2} a e^{-\hat{||x - \epsilon^{-1/2} z||}^2 / \sigma} \, d\hat{V}(\hat{z}) = O(\epsilon^2),
\]

where \( x - \gamma \sqrt{\epsilon^2} \) so \( dV(y) = \epsilon^{d/2} d\hat{V}(\hat{z}) \), where \( d\hat{V}(\hat{z}) \) is the volume form of the transformed manifold \( \hat{M} \), and the last inequality follows from extending the integral from the manifold to the entire ambient space, and the final equality follows from the exponential decay of the integral of the tail of a Gaussian distribution, since \( \epsilon^{\gamma - 1/2} \to \infty \) as \( \epsilon \to 0 \). This allows us to localize the integral in \( G_\epsilon \) to an \( \epsilon^2 \) neighborhood of \( x \) so that,

\[
G_\epsilon(f q)(x) = \epsilon^{-d/2} \int_M h \left( \frac{||x - y||^2}{\epsilon} \right) f(y) q(y) \, dV(y) = \epsilon^{-d/2} \int_{y \in M, ||x - y|| < \epsilon^2} h \left( \frac{||x - y||^2}{\epsilon} \right) f(y) q(y) \, dV(y) + O(\epsilon^2).
\]

The remainder of the proof proceeds by local asymptotic expansion of \( G_\epsilon \) following \([4]\) with no further modifications.

The above lemma shows that the central asymptotic expansion of \([4]\) extends trivially to non-compact manifolds by assuming an integrable sampling distribution \( q \) which is bounded above. Lemma 2.2 requires that \( f \in L^2(M, q) \) so the expansion is only valid for functions whose growth is controlled by the decay of the sampling \( q \). In particular, for unbounded manifolds such as \( \mathbb{R} \) (see Section 5) the eigenfunctions of the operator \( L_{\alpha, \beta} \) may have polynomial growth and thus we will typically assume \( q \) has exponential decay at infinity.

### 2.2. Left formulation of uniformly sampled data

Let \( \rho(x) \) be a positive function on the manifold and define the variable bandwidth Gaussian kernel, with bandwidth \( \rho \) to be

\[
K^\epsilon_\rho(x, y) = h \left( \frac{||x - y||^2}{\epsilon \rho(x)} \right)
\]

For simplicity we first assume uniform sampling on a compact manifold, since Section 2.1 shows that these expansions directly generalize by simply applying the operator \( G_\epsilon \) to the product \( f q \) under appropriate assumptions on \( f \) and \( q \). In
the Section 2.5 we will return to non-uniform sampling using this strategy. Under the uniform sampling assumption, the effect of the bandwidth function, \( \rho \), is to weight the Laplacian. To show this, we define the following change of variables, \( \hat{y} = \mathcal{F}(y) = \frac{\sqrt{\rho(x)}}{\sqrt{\rho(x)}} + x \). Then the integral operator,  
\[
G^L_\epsilon f(x) = \epsilon^{-d/2} \int_{\mathbb{R}^d} h\left( \frac{\|x - y\|^2}{\epsilon \rho(x)} \right) f(y) dV(y) 
= \epsilon^{-d/2} \rho(x)^{d/2} \int_{\mathcal{M}} h\left( \frac{\|x - y\|^2}{\epsilon} \right) f\left( \frac{\sqrt{\rho(x)}(\hat{y} - x) + x}{\sqrt{\rho(x)}} \right) d\hat{V}(\hat{y}) 
\]
where \( \left| \frac{\partial \phi}{\partial y} \right| = \rho^{d/2} \). We can now apply Lemma 2.1 to the integral expression with the function \( f(\hat{y}) = f\left( \frac{\sqrt{\rho(x)}(\hat{y} - x) + x}{\sqrt{\rho(x)}} \right) \) so that,  
\[
G^L_\epsilon f(x) = \rho(x)^{d/2} G^\epsilon_\rho f(x) = \rho(x)^{d/2} \left( m_0 f(x) + m_2 \epsilon(\omega(x)f(x)) + \Delta f(x) \right), 
\]
where \( m_0 \) and \( m_2 \) are constants determined by the shape function \( h \).  
\[
m_0 = \int_{\mathbb{R}^d} h\left( \|z\|^2 \right) dz, \quad m_2 = \frac{1}{2} \int_{\mathbb{R}^d} z_i^2 h\left( \|z\|^2 \right) dz, 
\]
which are the same expressions as for the fixed bandwidth kernel in Lemma 2.1. Note that the transformation \( \kappa(\hat{y}) = \sqrt{\rho(x)(\hat{y} - x) + x} \) inside the function \( f(\sqrt{\rho(x)}(\hat{y} - x)) \) corresponds to a change of metric in the tangent space \( T_x \mathcal{M} \). The map \( \kappa(\hat{y}) \) is a local diffeomorphism such that \( \kappa(x) = x \) and \( D\kappa(\hat{y}) = \frac{\sqrt{\rho(x)}}{\sqrt{\rho(x)}} I_{dx,dx} \). In this small neighborhood of \( x \), the Laplacian can be written locally as,  
\[
\Delta f(\hat{y}) = \frac{1}{\sqrt{|\nabla \rho|}} \partial_i \left( g^{ij} \sqrt{|\nabla \rho|} \partial_j f \kappa \right) \bigg|_{y=x} = \frac{1}{\sqrt{|\nabla \rho|}} \partial_i \left( g^{ij} \sqrt{|\nabla \rho|} \partial_j f \kappa \right) \bigg|_{y=x} = \frac{\rho(x)}{\sqrt{|\nabla \rho|}} \partial_i \left( g^{ij} \sqrt{|\nabla \rho|} \partial_j f \kappa \right) \bigg|_{y=x} = \rho(x) \Delta f(x). 
\]
Notice that the above result is simply two applications of the chain rule, combined with the fact that \( \kappa(x) = x \) and \( D\kappa(\hat{y}) = \frac{\sqrt{\rho(x)}}{\sqrt{\rho(x)}} I_{dx,dx} \). Combining (9) with (8) we have the following expansion for the \( G^L_\epsilon \) based on the variable bandwidth kernel,  
\[
G^L_\epsilon f(x) = \rho(x)^{d/2} \left( m_0 f(x) + m_2 \epsilon(\omega(x)f(x)) + \rho(x) \Delta f(x) \right) + O(\epsilon^2) 
= m_0 \rho(x)^{d/2} f(x) \left( 1 + \epsilon m \left( \omega(x) + \frac{\Delta f(x)}{f(x)} \right) \right) + O(\epsilon^2), 
\]
where \( m \equiv m_2 / m_0 \). We now apply a left-normalization, dividing by \( G^L_\epsilon 1(x) \) outside the operator so we have,  
\[
\frac{G^L_\epsilon f(x)}{G^L_\epsilon 1(x)} = f(x) + \epsilon m \rho(x) \Delta f(x) + O(\epsilon^2). 
\]
Finally, we can extract the order-\( \epsilon \) term, defining the operator \( L^L_\epsilon \) by,  
\[
L^L_\epsilon f(x) = \frac{1}{\epsilon m \rho(x)} \left( \frac{G^L_\epsilon f(x)}{G^L_\epsilon 1(x)} - f(x) \right) = \Delta f(x) + O(\epsilon). 
\]
In Figure 1, we numerically verify this asymptotic expansion with a simple example on a periodic domain. Of course we can only easily approximate the operator $G^2$ via a Monte-Carlo integral approximation,

$$
limit_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} K^2_{\epsilon}(x, x_i) f(x_i) = \frac{1}{\text{vol}(\mathcal{M})} \int_{\mathcal{M}} K^2_{\epsilon}(x, y) f(y) dV(y) = \frac{\epsilon^{d/2}}{\text{vol}(\mathcal{M})} G^2 \epsilon f(x),$$

which is only valid when the sampling density of $[x_i]$ is uniform on $\mathcal{M}$. We will return to the case of non-uniform sampling in Section 2.5. Note that the factor $\epsilon^{d/2}/\text{vol}(\mathcal{M})$ does not affect the operator $L^2$ due to the left-normalization. We now turn to the right formulation $K^R_{\epsilon}(x, y) = K^L_{\epsilon}(y, x)$ and we will emphasize the significant difference between these deceptively similar kernels.

2.3. Right formulation of uniformly sampled data

In this section we continue to assume uniform sampling and consider the right formulation of the variable bandwidth kernel,

$$K^R_{\epsilon}(x, y) = h \left( \frac{\|x - y\|^2}{\epsilon \rho(y)} \right)$$

which contains a variable bandwidth dependent only on $y$. As we will see, the dependence on $\rho(y)$ will not be the same as that of the variable bandwidth given by $\rho(x)$ considered in the previous section. In this case, it is not possible to simply change variables to eliminate the $\rho(y)$ term from the kernel, since any change of variables which involves both $y$ and $\rho(y)$ would not be explicitly invertible. Moreover, the Jacobian of such a change of variables would involve the gradient of the bandwidth function $\rho(y)$, which gives some insight into the difference between bandwidth functions which depend on $y$ rather than $x$.

To find an expansion of the following operator,

$$G^R_{\epsilon} f(x) = \epsilon^{-d/2} \int_{\mathcal{M}} h \left( \frac{\|x - y\|^2}{\epsilon \rho(y)} \right) f(y) dV(y),$$

we consider a weak formulation,

$$\langle g, G^R_{\epsilon} f \rangle = \epsilon^{-d/2} \int_{\mathcal{M}} \int_{\mathcal{M}} h \left( \frac{\|x - y\|^2}{\epsilon \rho(y)} \right) f(y) g(x) dV(y) dV(x),$$

for any arbitrarily smooth function $g$. We now simply exchange the order of integrations, and expand the inner integral using our previous result in (10),

$$\langle g, G^R_{\epsilon} f \rangle = \int_{\mathcal{M}} \epsilon^{-d/2} \left[ m_0 \rho(y) + \rho(y) \Delta g(y) \right] f(y) dV(y) + O(\epsilon^2)$$

where the last equality follows from the symmetry of $\Delta$, meaning $\langle f \rho^{d/2+1}, \Delta g \rangle = \langle \Delta (f \rho^{d/2+1}), g \rangle$. Recombining the two integrals, we summarize the above calculation as,

$$\langle g, G^R_{\epsilon} f \rangle = \int_{\mathcal{M}} g(y) \left[ m_0 \rho(y) \rho^{d/2} f(y) + \rho^{d/2} \Delta g(y) \right] dV(y) + O(\epsilon^2).$$

Finally, since $g(y)$ was arbitrary, we conclude that,

$$G^R_{\epsilon} f = m_0 \rho^{d/2} f + \rho^{d/2} \Delta (f \rho^{d/2+1}) + O(\epsilon^2).$$
Expanding \( \Delta(f^{d/2+1}) \) we find,

\[
\Delta(f^{d/2+1}) = f\Delta g^{d/2+1} + \rho^{d/2+1}\Delta f + (d+2)\rho^{d/2}\nabla\rho \cdot \nabla f,
\]

which allows us to write the operator expansion as,

\[
G_\rho^f f = \rho^{d/2}(m_0 f + em_2(\hat{\omega} f + \rho \Delta f + (d+2)\nabla\rho \cdot \nabla f)) + O(\epsilon^2),
\]

where \( \hat{\omega} = \omega - \rho^{-d/2}\Delta\rho^{d/2+1} \). If we apply the left normalization, dividing by \( G_\rho^f \) outside the operator, we find that,

\[
G_\rho^f f(x) \equiv \frac{G_\rho^f f(x)}{G_\rho^f 1(x)} = f(x) + em(\rho(x)\Delta f(x) + (d+2)\nabla\rho(x) \cdot \nabla f(x)) + O(\epsilon^2).
\]

Finally, we can extract the order-\( \epsilon \) term, defining the operator \( L_\rho^f \) by,

\[
L_\rho^f f \equiv \frac{1}{\text{emp}} \left( \frac{G_\rho^f f}{G_\rho^f 1} - f \right) = \Delta f + (d+2)\frac{\nabla\rho}{\rho} \cdot \nabla f + O(\epsilon).
\]

Note that when the variable bandwidth is a function of \( y \), the operator \( L_\rho^f \) takes the form of a Kolmogorov operator for diffusion in potential well given by \( U(x) = -(d+2)\log(\rho(x)) \). We verify this result in Figure 1. We now turn to the symmetric formulation with kernel \( K_\rho^f(x, y) \) which will require this result.

### 2.4. Symmetric bandwidth for uniformly sampled data

Continuing with our assumption of uniform sampling, we now return to the kernel,

\[
K_\rho^f(x, y) = h \left( \frac{||x - y||^2}{\epsilon\rho(x)\rho(y)} \right),
\]

and the associated operator,

\[
G_\epsilon^f f(x) = \epsilon^{-d/2} \int_M K_\rho^f(x, y)f(y) dV(y) = \epsilon^{-d/2} \int_M h \left( \frac{||x - y||^2}{\epsilon\rho(x)\rho(y)} \right) f(x) dV(y).
\]

In order to expand this expression, we will first change variables to eliminate the \( \rho(x) \) term and then we will apply the expansion in (11). Define the change of variables \( \tilde{x} = F(y) = x - \frac{\sqrt{\rho(x)}}{\sqrt{\rho(y)}} \) so that \( y = x - \sqrt{\rho(x)}(x - \tilde{x}) \) and

\[
G_\epsilon^f f(x) = \epsilon^{-d/2} \rho(x)^{d/2} \int_{F(M)} h \left( \frac{||\tilde{x} - \tilde{y}||^2}{\epsilon\rho(x)\rho(y)} \right) f(x - \sqrt{\rho(x)}(x - \tilde{x})) d\tilde{V}(\tilde{y}).
\]

Letting \( \hat{\rho}(\tilde{y}) = \rho(x - \sqrt{\rho(x)}(x - \tilde{y})) \) and \( \hat{f}(\tilde{y}) = f(x - \sqrt{\rho(x)}(x - \tilde{y})) \) we have,

\[
G_\epsilon^f f(x) = \epsilon^{-d/2} \rho(x)^{d/2} \int_{F(M)} h \left( \frac{||\tilde{x} - \tilde{y}||^2}{\epsilon\hat{\rho}(\tilde{y})} \right) \hat{f}(\tilde{y}) d\tilde{V}(\tilde{y}).
\]

Applying the expansion (11) from the previous section we have,

\[
G_\epsilon^f f = \rho^{d/2} \rho^{d/2} \left( m_0 \hat{f} + em_2(\hat{\omega} \hat{f} + \hat{\rho} \Delta \hat{f} + (d+2)\nabla\hat{\rho} \cdot \nabla \hat{f}) \right) + O(\epsilon^2).
\]

Note that \( \hat{\rho}(x) = \rho(x) \) and \( \hat{f}(x) = f(x) \) so that

\[
G_\epsilon^f f = \rho^{d} \left( m_0 f + em_2(\hat{\omega} f + \rho \Delta f + (d+2)\nabla\rho \cdot \nabla f) \right) + O(\epsilon^2).
\]

Furthermore, \( \nabla \hat{f}(x) = \nabla f(x + \sqrt{\rho(x)}(x - \tilde{y})) \bigg|_{\tilde{y}=x} = \sqrt{\rho(x)}\nabla f(x) \) and \( \nabla \hat{\rho}(x) = \sqrt{\rho(x)}\nabla \rho(x) \) and \( \Delta \hat{f}(x) = \rho(x)\Delta f(x) \) so we have,

\[
G_\epsilon^f f = \rho^{d} \left( m_0 f + em_2(\hat{\omega} f + \rho^2 \Delta f + (d+2)\rho \nabla\rho \cdot \nabla f) \right) + O(\epsilon^2).
\]
Applying left-normalization we find the operator $L^S$ given by,
\[
L^S f(x) \equiv \frac{1}{\epsilon m \rho(x)^2} \left( \frac{G^S_{q}(x)}{G^S_{1}(x)} - f(x) \right) = \Delta f + (d + 2) \frac{\nabla \rho}{\rho} \cdot \nabla f + O(\epsilon^2).
\]

We verify this formula in Figure 1 on a unit circle in $\mathbb{R}^2$ and a flat torus in $\mathbb{R}^4$.

**Remark.** We verify this formula in Figure 1 on a unit circle in $\mathbb{R}^2$ and a flat torus in $\mathbb{R}^4$. For general variable bandwidth with non-symmetric variable bandwidth Kernels,
\[
K^S_{U}(x,y) = \hat{h}\left( \frac{||x-y||^2}{\epsilon \rho_1(x)\rho_2(y)} \right),
\]
it is not difficult to check that under the uniform sampling assumption,
\[
L^U f(x) \equiv \frac{1}{\epsilon m \rho_1(x)\rho_2(x)} \left( \frac{G^U_{q}(x)}{G^U_{1}(x)} - f(x) \right) = \Delta f + (d + 2) \frac{\nabla \rho_2}{\rho_2} \cdot \nabla f + O(\epsilon^2),
\]
where $G^U_{q} \equiv \epsilon^{-d/2} \int M K^U_{q}(x,y) f(y) dV(y)$.

### 2.5. Symmetric bandwidth for non-uniformly sampled data

Using the expansion of the symmetric variable bandwidth kernel from the previous section, we can now extend the result to the case of non-uniform sampling following the strategy of Coifman and Lafon in [4]. Assume a positive sampling measure with density function $q(x)$ on $M$, then when we compute Monte-Carlo approximations of kernel operators we will find,
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} K^S_{\epsilon}(x,x_i) f(x_i) = \int_{M} K^S_{\epsilon}(x,y) f(y) dV(y) = \epsilon^{d/2} \int_{M} G^S_{\epsilon}(f q)(x).
\]

The previous equation implies that the direct application of our kernel $K^S_{\epsilon}$ will be biased by $q(x)$. Thus we define the biased operator,
\[
G^S_{\epsilon,q}(f) \equiv G^S_{\epsilon}(f q).
\]

In order to remove the bias we first estimate $q(x)$ in the sense of a kernel density estimate by setting $f(x) = 1$. Applying the result (12) from the previous section and introducing the abbreviation $L^S(f) = \rho^2 \Delta f + (d + 2) \rho \nabla \rho \cdot \nabla f$ we have,
\[
G^S_{\epsilon,q}(1) = G^S_{\epsilon}(q) = \rho d \left( m_0 q + \epsilon m_2 (\omega q + \rho^2 \Delta q) + (d + 2) \rho \nabla \rho \cdot \nabla q \right) + O(\epsilon^2)
\]
\[
= m_0 \rho^d q \left( 1 + \epsilon m \left( \omega + L^S q \right) \right) + O(\epsilon^2)
\]
\[
\begin{align*}
G^S_{\epsilon,q}(1)^\alpha & = \left( m_0 \rho^d q \right)^\alpha \left( 1 + \epsilon m \left( \omega + L^S q \right) \right) + O(\epsilon^2),
\end{align*}
\]

As in the standard Diffusion Map formulation [4], we introduce the “de-biasing” parameter $\alpha$ and note that,
\[
G^S_{\epsilon,q,q}(1)^\alpha = \left( m_0 \rho^d q \right)^\alpha \left( 1 + \epsilon m \left( \omega + L^S q \right) \right) + O(\epsilon^2),
\]

In particular, the sampling bias in the kernel operator is removed through a right normalization,
\[
G^S_{\epsilon,q,q}(f) \equiv G^S_{\epsilon,q} \left( f \frac{\rho^d q}{G^S_{\epsilon,q}(1)^\alpha} \right).
\]

Note that while this normalization appears to make the kernel non-symmetric, in Section 4 we present a numerical technique which will allow us to maintain the symmetry of the kernel matrix for the purpose of finding eigenvalues. Applying the result (12) from the previous section we have,
\[
G^S_{\epsilon,q,q}(f) = G^S_{\epsilon,q} \left( \frac{f \rho^d q}{G^S_{\epsilon,q}(1)^\alpha} \right)
\]
\[
= m_0 \rho^d q \left( 1 + \epsilon m \left( \omega + L^S (f \rho^d q G^S_{\epsilon,q}(1)^\alpha) \right) \right) + O(\epsilon^2)
\]
\[
= m_0 \rho^d q \left( 1 - \epsilon m \left( \omega + L^S q \right) \right) \left( 1 + \epsilon m \left( \omega + L^S (f q(m_0 q)^{-\alpha}) \right) \right) + O(\epsilon^2)
\]
\[
= f \rho^d q \left( 1 - \epsilon m \left( \omega + L^S q \right) \right) \left( 1 + \epsilon m \left( (1 - \alpha) \omega - \alpha L^S q + L^S (f q(m_0 q)^{-\alpha}) \right) \right) + O(\epsilon^2)
\]
\[
= f \rho^d q \left( 1 + \epsilon m \left( 1 - \alpha \omega \right) - \alpha L^S q + L^S (f q(m_0 q)^{-\alpha}) \right) + O(\epsilon^2)
\]
Figure 1: Operators $L_t^f$ (top), $L_\Delta f$ (middle), and $L_\rho f$ (bottom) with variable bandwidth $\rho(\theta) = \exp(\cos(\theta))$ applied to $f(\theta) = \sin(\theta)$ (left column) and $f(\theta, \phi) = \sin(\theta)$ (right column). Functions and operators are constructed on a uniform grid of 3000 points on the unit circle (left) and a uniform grid of 62,500 points on a flat torus in $\mathbb{R}^4$ (right). Each operator is constructed for $\epsilon = 0.1$ (blue), 0.01 (green), and 0.001 (red). Left, top: $\Delta f = -\sin(\theta)$ (grey) compared to $L_t^f f$. Left, middle: $\Delta f + (d + 2) \nabla_\rho \cdot \nabla f = -\sin(\theta) - 3 \sin(\theta) \cos(\theta)$ (grey) compared to $L_\rho f$. Left, bottom: $\Delta f + (d + 2) \nabla_\rho \cdot \nabla f = -\sin(\theta) - 3 \sin(\theta) \cos(\theta)$ (grey) compared to $L_\rho f$. Right, top: $\Delta f = -\sin(\theta)$ (grey) compared to $L_t^f f$. Right, middle: $\Delta f + (d + 2) \nabla_\rho \cdot \nabla f = -\sin(\theta) - 4 \sin(\theta) \cos(\theta)$ (grey) compared to $L_\rho f$. Right, bottom: $\Delta f + (d + 2) \nabla_\rho \cdot \nabla f = -\sin(\theta) - 4 \sin(\theta) \cos(\theta)$ (grey) compared to $L_\rho f$. Note that on the torus we use a sparse matrix construction of the operators where only the 500 nearest neighbors of each point are allowed nonzero entries, this degrades the result for large $\epsilon$ but has no effect as $\epsilon$ becomes small due to the exponential decay of the kernel.
Now applying left normalization we find,

\[
\frac{G^S_{ε,α,β}(f)}{G^S_{ε,α,β}(1)} = \frac{f \rho^d(m_0q)^{1-α}}{ρ^d(m_0q)^{1-α}(1 + em((1 - α)ω - α L^S q + L^S(q(m_0q)^{-α})) + O(ε^2))}
\]

\[
= f \left(1 + em \left(L^S(fq(m_0q)^{-α}) - L^S(q(m_0q)^{-α})\right)\right) + O(ε^2)
\]

Extracting the order-ε term we have the operator,

\[
L^S_{ε,α}f(x) \equiv \frac{1}{\text{emp}(x)^2}\left(\frac{G^S_{ε,α,β}f(x)}{G^S_{ε,α,β},1(x)} - f(x)\right) = \frac{f}{ρ^2} \left(L^S(fq(m_0q)^{-α}) - L^S(q(m_0q)^{-α})\right) + O(ε)
\]

\[
= \frac{f}{ρ^2} \left(\frac{Δ(fg)}{g} + (d + 2)ρ^2 g \cdot \frac{ν(fg)}{g} - ρ^2 \frac{Δg}{g} - (d + 2)ρ^2 \frac{νg}{g}\right) + O(ε)
\]

\[
= Δf + 2νf \cdot \frac{νg}{g} + (d + 2)νf \cdot \frac{νρ}{ρ} + O(ε)
\]

where \(g \equiv m_0^{-α}q^{1-α}\) is introduced for convenience. Note that \(\frac{Δ}{g} = (1-α)\frac{Δg}{g}\) so we can simplify the previous expression to,

\[
L^S_{ε,α}f(x) = Δf + 2(1-α)νf \cdot \frac{νq}{q} + (d + 2)νf \cdot \frac{νρ}{ρ} + O(ε)
\]

Note that (17) shows how the variable bandwidth function \(ρ\) affects the operator defined by the kernel. When \(ρ = 1\) is constant, we recover the result of [4], namely a gradient flow with potential function \(U = -2(1-α)log q\) defined by the sampling density \(q\). The formula (17) reveals that we can use a variable bandwidth kernel to approximate the generator of a gradient flow for an arbitrary potential function \(U\) by choosing bandwidth function \(ρ = e^{-U/(d+2)}\) so that \((d + 2)\frac{Δg}{g} = -∇U\) is the vector field defined by the gradient of the potential function \(U\). Setting \(α = 1\) as in [4], we remove the effect of the sampling density \(q\) on the operator, and recover the desired generator \(Δf - ∇U \cdot ∇f\).

Finally, if we make the choice \(ρ = q^d\), we find,

\[
L^S_{ε,α,β}f(x) = Δf + c_1 νf \cdot \frac{νq}{q} + O(ε)
\]

where \(c_1 = 2 - 2α + dβ + 2β\). In Figure 2, we numerically verify the expansion in (18) on a circle sampled according to the distribution \(q(θ) = \exp(\cos(θ))\). Since \(d = 1\), setting \(β = -1/2\) and \(α = 1/4\) we find \(c_1 = 0\) and we recover the Laplacian on the circle. Setting \(β = -1/2\) and \(α = 1/4\) we find \(c_1 = 1\) which yields the Kolmogorov operator for the potential \(U(θ) = -log(q(θ))\) on the circle.

In practical applications the sampling density, \(q\), will usually not be known, however we can always use any kernel to estimate the sampling density, see [13, 12, 16, 14] as well as the numerical details in Section 4. Of course, this means that we will actually have \(ρ = q^d + O(ε)\) for \(N\) sufficiently large). While this approximation will affect the expansion of the kernel \(K^S_ε\), it is easy to see that all of these effects are canceled by the left-normalization. Since the estimate in (18) is already order-ε this result is not affected by order-ε errors in the sampling density estimate which is used for the bandwidth function.

3. Convergence Rates for Discrete Operators

The goal of this section is to analyze the the accuracy of the discrete estimates of the continuous kernel operators defined above. Here, we follow the analysis of Singer [15] and generalize the error estimates to the case of variable bandwidth kernels and non-uniform sampling. Let \(\{x_j\}_{j=1}^N\) be independently sampled according to the density \(q(x)\) on
the manifold $M \subset \mathbb{R}^n$ (note that $M$ is any Riemannian manifold and is not assumed to be compact). For fixed $x = x_i$ from the data set, define the random variables,

$$F_i(x_j) = \frac{K^S(x_i,x_j)f(x_j)}{(e^{-d/2}q^S(x_j))^\alpha}, \quad G_i(x_j) = \frac{K^S(x_i,x_j)}{(e^{-d/2}q^S(x_j))^\alpha},$$

where $q^S(x_j) = \frac{1}{N} \sum_{j'} K^S(x_j,x_{j'})/\rho(x_{j'})^d$. The functionals $F_i$ and $G_i$ are used in the numerical algorithm to approximate the operator,

$$L_{x,\alpha,\beta}^S f(x_i) = \frac{1}{\text{emp}(x_i)^2} \left( \frac{\mathbb{E}[F_i]}{\mathbb{E}[G_i]} - f(x_i) \right) \approx \frac{1}{\text{emp}(x_i)^2} \left( \frac{\sum_j F_i(x_j)}{\sum_j G_i(x_j)} - f(x_i) \right),$$

where the continuous expectations are defined as,

$$\mathbb{E}[F_i] \equiv \int_M F_i(y)q(y)\,dV(y), \quad \mathbb{E}[G_i] \equiv \int_M G_i(y)q(y)\,dV(y)$$

so that the continuous operator $L_{x,\alpha,\beta}^S$ agrees with the previous theory in Section 2.5. Notice that the factor $\frac{e^{d/2}}{N}$ does not need to be known or included in the actual algorithm, since ultimately we will be interested in the ratio $\sum_j F_i(x_j)/\sum_j G_i(x_j)$ and the factor cancels exactly. Similarly, while the algorithm and statement of Theorem 1 divide each functional by $q^S(x_j)^\alpha$, this factor cancels in $L_{x,\alpha,\beta}^S$ since the expectations are taken with respect to $x_j$. Finally, since the density, domain, and number of sample points are the same, the normalization factors Monte-Carlo summations, $\sum_j F_i(x_j)$ and $\sum_j G_i(x_j)$, are identical and are therefore left out.

In the subsequent sections we will find the error in replacing the continuous expectations with the discrete sums. If we consider the approximation above to be an estimator for $L_{x,\alpha,\beta}^S f(x_i)$ then the bias of the estimator is,

$$\mathbb{E} \left[ \frac{1}{\text{emp}(x_i)^2} \left( \frac{\sum_j F_i(x_j)}{\sum_j G_i(x_j)} - \frac{\mathbb{E}[F_i]}{\mathbb{E}[G_i]} \right) \right].$$

However, since this expectation is difficult to evaluate, we instead follow the analysis of Singer in [15], which bounds the probability of a large bias error by estimating,

$$P \left( \frac{1}{\text{emp}(x_i)^2} \left( \frac{\sum_j F_i(x_j)}{\sum_j G_i(x_j)} - \frac{\mathbb{E}[F_i]}{\mathbb{E}[G_i]} \right) > \alpha \right).$$

In order to estimate this error, we first need to control the error of the denominators in the functionals $F_i$ and $G_i$. 

Figure 2: Operators with variable bandwidth $\rho(x)q(y) = q(x)q(y)\beta$ with $\beta = -1/2$ are applied to $f(x) = \sin(x)$ where $x$ parameterizes a unit circle in the plane $\mathbb{R}^2$. Functions and operators are constructed on a set of 8000 points on the circle, sampled from the density $q(x) = \exp(\cos(x))$. Each operator is constructed for $\epsilon = 0.1$ (blue), 0.01 (green), and 0.005 (red). Left: $\Delta f(x) = -\sin(x)$ (grey) compared to $L_{x,\alpha,\beta}^S f$ with $\alpha = 1/4, \beta = -1/2$, note that this is the Laplacian operator. Right: $\Delta f(x) + \frac{\epsilon}{\rho(x)} \nabla f(x) = -\sin(x) - \sin(x)\cos(x)$ (grey) compared to $L_{x,\alpha,\beta}^S f$ with $\alpha = 1/4, \beta = -1/2$, notice that this is the backward Kolmogorov operator for gradient flow with potential $U = -\log(q) = -\cos(x)$.
3.1. Sampling error in the renormalization factor

To analyze the denominator terms of $F_i$ and $G_i$, let’s define $H_j(x_i) \equiv e^{-d/2}K^5_j(x_j,x_i)/\rho(x_j)^d$. From Section 2.4 we have,

$$\mathbb{E}[H_j(x_i)] = \rho(x_j)^{-d}e^{-d/2} \int_M K^5_j(x_j,x_i)q(y) dV(y) = \rho(x_j)^{-d}G^5_j(1) = m_0q(x_j) \left(1 + \epsilon m(x_j) + L^5 q(x_j))\right) + O(\epsilon^2),$$

where we use the expansion (14).

We first note that in order for the random variables $H_j(x_i)$ to be identically distributed, we must neglect the term $l = j$. This term is typically included in the implementation of the algorithm, however, the error made by neglecting it is estimated from the expansion,

$$\frac{1}{N} \sum_{l} H_j(x_i) = \frac{1}{N-1+1} \sum_{l \neq j} H_j(x_i) + \frac{e^{-d/2}}{N} = \frac{(N-1)^{-1}}{1+(N-1)^{-1}} \sum_{l \neq j} H_j(x_i) + \frac{e^{-d/2}}{N}$$

which shows that the error is $\frac{1}{N} \sum_{l} H_j(x_i) - \frac{1}{N-1} \sum_{l \neq j} H_j(x_i) = O(N^{-1}e^{-d/2})$. In the remaining of this section, we will use this error bound to replace the summation over all $l$ with the summation over $l \neq j$.

We now analyze the error between the discrete Monte-Carlo approximation, $\frac{1}{N^{-1}} \sum_{l} H_j(x_i)$, and the continuous expectation, $\mathbb{E}[H_j]$. Letting, $Y_i = H_j(x_i) - \mathbb{E}[H_j]$, we note that $\mathbb{E}[Y_i] = 0$ and for $l \neq j$,

$$\text{var}(Y_i) = \mathbb{E}[Y_i^2] = \mathbb{E}[H_j(x_i)^2] - \mathbb{E}[H_j]^2$$

$$= \tilde{m}_0\epsilon^{-d/2}\rho(x_j)^{-d}q(x_j) - m_0^2q(x_j)^2 + O(\epsilon^{-d/2}),$$

$$= \hat{m}_0\epsilon^{-d/2}\rho(x_j)^{-d}q(x_j) + O(1),$$

where $\hat{m}_0 \equiv \int_{\mathbb{R}^d} h(\|z\|^2)dz$. So by the Chernoff inequality we have, for $a$ sufficiently small,

$$P \left( \frac{1}{N-1} \sum_{l \neq j} H_j(x_i) - (N-1)\mathbb{E}[H_j] > a \right) = P \left( \sum_{l \neq j} Y_i > a(N-1) \right) \leq 2 \exp \left( \frac{-a^2(N-1)}{4\hat{m}_0\epsilon^{-d/2}\rho(x_j)^{-d}q(x_j)} \right). \quad (19)$$

Note the crucial fact that $q(x_j)$ appears in the denominator, so that as $\epsilon \to 0$ the probability of error in the estimate decays.

Recall that our goal is to expand the ratio $\mathbb{E}[F_i]/\mathbb{E}[G_i]$ up to order-$\epsilon^2$. Thus, we require the denominators of $F_i$ and $G_i$, given by $\epsilon^{-d/2}q^5_i(x_i) = \frac{1}{N-1} \sum_{l \neq j} K^5_j(x_j,x_i)/\rho(x_j)^d$ to agree with the continuous limits $\mathbb{E}[H_j]$ up to order-$\epsilon^2$. Thus we require,

$$\left| \epsilon^{-d/2}q^5_i(x_i) - \mathbb{E}[H_j] \right| = \left| \frac{\epsilon^{-d/2}}{N} \sum_{l \neq j} K^5_j(x_j,x_i)/\rho(x_j)^d - \mathbb{E}[H_j] \right|$$

$$= \left| \frac{\epsilon^{-d/2}}{N} \sum_{l \neq j} K^5_j(x_j,x_i)/\rho(x_j)^d - \frac{\mathbb{E}[H_j]}{N} \right| + O \left( \frac{\epsilon^{-d/2}}{N} \right) = O \left( 2^{d/2} \epsilon^{-d/2} \right)$$

with high probability. Notice, that balancing the two error terms requires $\epsilon = O(N^{-1/2+d/2})$. If we assume that $a = O(\epsilon^2)$ in (19), then we can write $a = \tilde{a}\epsilon^2$ where $\tilde{a} = O(1)$ and we will achieve the desired accuracy with high probability when the exponent of the Chernoff inequality, $\tilde{a}^2N\epsilon^{d/2}\rho(x_j)^d/q(x_j) = \tilde{a}^2N\epsilon^{4-d/2}\rho(x_j)^d/q(x_j)$, is large. In other words, for $\rho = \rho^d$, when

$$\frac{q(x_j)^{(1-2d)/2}}{N^{1/2}\epsilon^{2+d/2}} = O(1), \quad (20)$$

13
we attain the desired accuracy with high probability. Notice, that for $q = O(1)$ this requires $\epsilon = O(N^{-1/(4+d/2)})$ which dominates the previous requirement, $\epsilon = O(N^{-1/(2+d/2)})$. This shows that the error of order $O(\epsilon^{-d/2}/N)$ from neglecting the diagonal term is negligible compared to the error due to the variance of $Y_i$.

3.2. Bounding the statistical bias in the discrete estimate

Using the above estimate, we can now consider the case where the summations in the denominators can be replaced by the continuous expectations, so that when $\frac{q(t, x, g)}{\mathcal{Q}(x, t)} = O(1)$, we have,

$$F_i(x_i) = \frac{K^g_i(x_i, x_j)f(x_j)}{[E[H_i] + O(\epsilon^2)]} = \frac{K^g_i(x_i, x_j)f(x_j)}{m_0^g q(x_j)^g} (1 - \alpha \epsilon \omega(x_j) - L^\alpha q(x_j)) + O(\epsilon^2),$$

$$G_i(x_i) = \frac{K^g_i(x_i, x_j)}{[E[H_i] + O(\epsilon^2)]} = \frac{K^g_i(x_i, x_j)}{m_0^g q(x_j)^g} (1 - \alpha \epsilon \omega(x_j) - L^\alpha q(x_j)) + O(\epsilon^2).$$

From the expansion in (15), we deduce,

$$\mathbb{E}[F_i] = e^{d/2} G_{x,q}(\alpha) f = e^{d/2} f \rho^d (m_0 q)^{1/\alpha} (1 + \epsilon m((1 - \alpha) \omega - \alpha L^\alpha q + L^\alpha (\epsilon q(m_0 q)^{-\alpha}))) + O(\epsilon^{2+d/2}),$$

$$\mathbb{E}[G_i] = e^{d/2} G_{x,q}(\alpha) (1) = e^{d/2} f \rho^d (m_0 q)^{1/\alpha} (1 + \epsilon m((1 - \alpha) \omega - \alpha L^\alpha q + L^\alpha (\epsilon q(m_0 q)^{-\alpha}))) + O(\epsilon^{2+d/2}).$$

Therefore, we can deduce

$$\mathbb{E}[F_i^2] = e^{d/2} f^2 \rho^d q^{1/\alpha} m_0 (1 + \epsilon \omega(m - 2 \alpha m) - 2 \epsilon m L^\alpha q + \epsilon m L^\alpha (\epsilon q(m_0 q)^{-\alpha}))) + O(\epsilon^{2+d/2}),$$

$$\mathbb{E}[G_i^2] = e^{d/2} f^2 \rho^d q^{1/\alpha} m_0 (1 + \epsilon \omega(m - 2 \alpha m) - 2 \epsilon m L^\alpha q + \epsilon m L^\alpha (\epsilon q(m_0 q)^{-\alpha}))) + O(\epsilon^{2+d/2}),$$

$$\mathbb{E}[F_i G_i] = e^{d/2} f^2 \rho^d q^{1/\alpha} m_0 (1 + \epsilon \omega(m - 2 \alpha m) - 2 \epsilon m L^\alpha q + \epsilon m L^\alpha (\epsilon q(m_0 q)^{-\alpha}))) + O(\epsilon^{2+d/2}),$$

where $\hat{m}_0 \equiv \int_{S^d} h(\|z\|^2)^2 dz$ and $\hat{m}_2 \equiv \int_{S^d} \epsilon^2 h(\|z\|^2)^4 dz$.

Following the analysis of [15] we want to compute,

$$P\left(\frac{1}{\text{emp}(\epsilon^2)} \sum_{j \neq i} F_j(x_i) > a \right) = P\left(\sum_{j \neq i} Y_j > a(N-1)\mathbb{E}[G_i]^2 \text{emp}(\epsilon^2) \right)$$

where $Y_j = \mathbb{E}[G_i F_j(x_j) - \mathbb{E}[F_j G_i(x_j) + \epsilon \text{emp}(\epsilon^2) \mathbb{E}[G_i] \mathbb{E}[G_i] - G_i(x_j))$. Note that $\mathbb{E}[Y_j] = 0$ and the variance is given by,

$$\mathbb{E}[Y_j^2] = \mathbb{E}[G_i]^2 \mathbb{E}[F_j^2] + \mathbb{E}[F_j]^2 \mathbb{E}[G_i^2] - 2 \mathbb{E}[G_i] \mathbb{E}[F_j] \mathbb{E}[F_i G_i] + O(\epsilon \text{emp})$$

It is easy to see that the order-1 term in the variance is zero, computing the order-1 term we note that the $\omega$ terms and $L^\alpha q$ terms also cancel. Letting $\mathcal{H} f = f L^\alpha f$ we have,

$$\mathbb{E}[Y_j^2] = \epsilon^{1+3d/2} \rho^{3d} m_0^{2-6d} \hat{m}_2 \left(q^{2-2\alpha} \mathcal{H}(f q^{1-2\alpha}) + f^2 q^{2-2\alpha} \mathcal{H}(q^{1-2\alpha}) - 2 f q^{2-2\alpha} \mathcal{H}(f q^{1-2\alpha}) \right) + O(\epsilon \text{emp} + \epsilon^{2+d/2}).$$

Note that $\mathcal{H} f = f L^\alpha f = \rho^2 L f + (d + 2) \rho \nabla f \cdot \nabla f$, so that for arbitrary $f, g$ we have,

$$L^\alpha (f g) = \rho^2 L f g + (d + 2) \rho \nabla f \cdot \nabla (f g)$$

$$ = \rho^2 (f L g + g L f + 2 \nabla f \cdot \nabla g) + (d + 2) \rho \nabla f \cdot (f \nabla g + g \nabla f)$$

$$ = f L g + g L f + 2 \rho \nabla f \cdot \nabla g.$$

Using the above property, we can simplify the variance as,

$$\mathbb{E}[Y_j^2] = 2 \epsilon^{1+3d/2} \rho^{3d} m_0^{2-6d} \hat{m}_2 \left(q^{2-2\alpha} \mathcal{H}(f q^{1-2\alpha}) - \rho^2 f q^{2-2\alpha} \mathcal{H}(q^{1-2\alpha}) \right) + O(\epsilon \text{emp} + \epsilon^{2+d/2})$$

$$ = 2 \epsilon^{1+3d/2} \rho^{3d} m_0^{2-6d} \hat{m}_2 \left(q^{2-2\alpha} \mathcal{H}(f q^{1-2\alpha}) + O(\epsilon \text{emp} + \epsilon^{2+d/2}) \right)$$

$$ = 2 \rho^{3d} m_0^{2-6d} \hat{m}_2 \epsilon^{1+3d/2} \left(q^{2-4d} \rho^{3+3d} \|\nabla f\|^2 + O(\epsilon \text{emp} + \epsilon^{2+d/2}) \right)$$
Finally, by the Chernoff bound we have,

\[
P\left( \sum_{j=1}^{m} Y_j > a(N-1)\mathbb{E}[G_j] + \epsilon \text{emp}^2 \right) \leq 2 \exp \left( \frac{-\alpha^2 (N-1)^2 \mathbb{E}[G_j]^2}{4(N-1) \text{var}(Y_j)} \right) = 2 \exp \left( \frac{-\alpha^2 (N-1)m^2 \epsilon^4 e^{-2d/\alpha} (\rho^2 + e^{2d/2})}{8m_0^2 \epsilon \rho^2 \mathbb{E}[\nabla f]^2 + O(\epsilon)} \right) = 2 \exp \left( \frac{-\alpha^2 (N-1)c_1 \rho^4 \epsilon^4 e^{-2d/\alpha} (\rho^2 + e^{2d/2})}{4e^{-1-d/2} \mathbb{E}[\nabla f]^2} \right),
\]

where \( c = m^2 m_0^2 e^{-6d} / (8m_0^2) \). By choosing bandwidth \( \rho = q^\beta \) the bound becomes,

\[
P\left( \frac{1}{\text{emp}(x) \mathbb{E}[G_j]} \left( \sum_{j=1}^{m} G_j(x_i) \right) | \mathbb{E}[G_j] > a \right) \leq 2 \exp \left( \frac{-\alpha^2 (N-1)m^2 q^4 \epsilon^4 e^{-2d/\alpha} (\rho^2 + e^{2d/2})}{4e^{-1-d/2} \mathbb{E}[\nabla f]^2} \right)
\]

so when the exponent is \( O(1) \) we can solve for \( \alpha \) to find the expected magnitude of errors to be,

\[
O\left( \frac{\|\nabla f\|^q q^{-c_2}}{N^{1/2} e^{1/2d/4}} \right)
\]

(23)

where \( c_2 = 1/2 - 2\alpha + 2\alpha a + \beta \). Notice that the error will be proportional to \( q^{-c_2} \) so for \( c_2 > 0 \) the errors may be unbounded as \( q \to 0 \).

Combining the error estimates (18), (20), and (23) we have,

\[
L_{s,a}^f(x_i) = \frac{1}{\text{emp}(x_i)^2} \left( \sum_{j=1}^{m} F_j(x_i) - f(x_i) \right) = \Delta f(x_i) + c_1 \nabla f(x_i) \cdot \overline{\nabla q(x_i)} / \sqrt{q(x_i)} + \mathcal{O} \left( \frac{\|\nabla f(x_i)\|^q q(x_i)^{-c_2}}{\sqrt{N} e^{1/2d/4}} \right),
\]

which proves Theorem 1.

4. Details of the numerical implementation

Given a data set \( \{x_i\}_{i=1}^{N} \subset \mathbb{R}^n \) sampled independently from a density \( q(x) \) on a \( d \)-dimensional Riemannian manifold \( M \subset \mathbb{R}^n \), the algorithm of this section will produce an \( N \times N \) sparse matrix which approximates the Kolmogorov operator,

\[
L_{s,a}^f f(x) = \Delta f + c_1 \nabla f \cdot \overline{\nabla q} / \sqrt{q},
\]

where the constant \( c_1 \) is determined by \( a \) and \( \beta \) as in Theorem 1. For example, assume that the data is generated by Brownian motion on a manifold in a potential \( U(x) \), that is,

\[
dx = -c_1 \nabla U(x) dt + \sqrt{a} dW_t,
\]

(24)

where \( W_t \) is a Brownian motion on the manifold \( M \) and \( U : M \to \mathbb{R} \) is smooth potential. The invariant measure of this system is given by \( q(x) = \exp(-c_1 U(x)) \) and we will assume that the data are independently sampled from this distribution. Let \( \Delta \) be the Laplacian (with negative eigenvalues) on \( M \), the generator of the stochastic process (24) is the backward Kolmogorov operator \( L_{s,a}^f \). Typically we will be interested in the cases \( c_1 = 0 \), which approximates the Laplacian (the generator for Brownian motion on \( M \)), and \( c_1 = 1 \) which approximates the generator of the stochastically forced gradient flow in (24).

In order to make use of the result in Theorem 1, we require the bandwidth function \( \rho \) to be a power of the sampling density, \( \rho = q^\beta + \mathcal{O}(\epsilon) \). While it is possible to use a fixed bandwidth kernel to estimate \( q \) up to order-\( \epsilon \), the results of section 3.1 suggest that we cannot take \( \epsilon \) very small unless \( N \) is large. The standard theory of variable bandwidth kernel density estimation [13, 12, 16, 14] offers multiple competitive algorithms, however for simplicity we will use an ad hoc method based on the distance to the nearest neighbors. To estimate \( q \) for the purposes of defining \( \rho \), we first
We must make sure to scale the eigenvectors appropriately. Since the eigenvectors approximate the eigenfunctions \( S \), we have,

\[
\mathbf{U} = \hat{\mathbf{L}} \mathbf{1}
\]

In order to find the eigenvectors of \( L \), we first find the eigen-decomposition of \( \hat{L} = \hat{U} \Lambda \hat{U}^\top \), and then note that setting \( U = S^{-1} \hat{U} \) we have,

\[
LU = S^{-1} \hat{L} S U = S^{-1} \hat{L} S S^{-1} \hat{U} = S^{-1} \hat{L} \hat{U} = S^{-1} \Lambda \hat{U} = \Lambda U,
\]

since \( S \) is diagonal. Thus, the columns of \( U \) are the desired eigenvectors of \( L \) with associated eigenvalues given by \( \Lambda \).

In order to make a comparison between the true eigenfunctions and the eigenvectors which approximate them, we must make sure to scale the eigenvectors appropriately. Since the eigenvectors approximate the eigenfunctions evaluated on the data set itself, they are sampled according to the density \( q \). For an eigenvector \( \phi = (\phi_1, ..., \phi_N) \), define an ad hoc bandwidth function \( \rho_0(x) = \left( \frac{1}{k_0} \sum_{j=1}^{k_0} \| x_i - x_{i(j)} \|^2 \right)^{1/2} \), where \( I(i,j) \) is the index of the \( j \)-th nearest neighbor of \( x_i \) from the data set (note that we leave out the nearest neighbor \( I(i,1) \), which is always the point \( x_i \) itself).

In the numerical examples in the next section we use \( k_0 = 8 \) nearest neighbors and we found that the results are not very sensitive to the choice of \( k_0 \leq 64 \). We then define \( \rho_0 \equiv \frac{1}{k_0} \sum_{i=1}^{k_0} \rho_0(x_i) \) and \( \tilde{\rho}_0 = \rho_0/\epsilon_0^{1/2} \), so that \( \tilde{\rho}_0 = O(1) \) and use a symmetric kernel with bandwidth \( \rho_0 \) to estimate the density as,

\[
q_0(x_i) = \frac{(2\pi)^{d/2}}{\tilde{\rho}_0(x_i) d N} \sum_{j=1}^{N} \exp \left( -\frac{\| x_i - x_j \|^2}{2\tilde{\rho}_0(x_i) \rho_0(x_j)} \right)
\]

\[
= \frac{(2\pi\epsilon_0)^{-d/2}}{\rho_0(x_i) d N} \sum_{j=1}^{N} \exp \left( -\frac{\| x_i - x_j \|^2}{2\epsilon_0 \rho_0(x_i) \rho_0(x_j)} \right) = q(x_i) + O \left( \epsilon_0, \frac{\sqrt{q(x_i)}}{N^{1/2} \epsilon_0 \rho_0(x_i)^{d/2}} \right),
\]

where the estimate follows from (18) and (20) with high probability. We can then use \( \rho \equiv q_0^2 = q^2 \) as the bandwidth function in the kernel \( K^S \) below. Balancing the two error terms in (25) we find, \( \epsilon_0 = O(\epsilon^{0.5/(d+4)}) \). Notice that \( \epsilon_0 \) is significantly smaller than \( \epsilon \) as required by balancing the error terms in Theorem 1, so that \( \rho = q^2 + O(\epsilon) \) as required in Theorem 1.

Using the bandwidth function \( \rho \) estimated as above, we now evaluate the kernel \( K^S \) on all pairs from the data set, and normalize following Theorem 1 to form \( L^S_{\epsilon,a,b} \), as,

\[
K^S_{\epsilon,a,b}(x_i,x_j) = \frac{\rho(x_i)}{q^2(x_i)} K^S_{\epsilon,a,b}(x_i,x_j)
\]

\[
L^S_{\epsilon,a,b}(x_i,x_j) = \frac{K^S_{\epsilon,a,b}(x_i,x_j)}{\epsilon \rho(x_i)}
\]

Note that the kernel \( K^S_{\epsilon,a,b} \) is symmetric, however, due to the left normalization, the kernel \( \tilde{K}^S_{\epsilon,a,b} \) is not symmetric and the need to normalize by \( \rho(x_i)^2 \) further degrades the symmetry in \( L^S_{\epsilon,a,b} \). Since we are interested in the eigenvalues and eigenvectors of \( L^S_{\epsilon,a,b} \), we instead construct a symmetric matrix which is given by conjugation of \( L^S_{\epsilon,a,b} \).

Let \( D_\rho = q_\rho^2(x_i) \) and \( P_\rho = \rho(x_i) \) be diagonal \( N \times N \) matrices and define the symmetric matrix, \( K_{\epsilon,a,b} = K^S_{\epsilon,a,b}(x_i,x_j) \). Let \( L_{\epsilon,a,b} = L^S_{\epsilon,a,b}(x_i,x_j) \) be the desired normalized Laplacian matrix. Note that, \( L = P^{-2}(D^{-1} K - I)/\epsilon \) and since \( P \) and \( D \) are diagonal, we can form the conjugation of \( L \) by the diagonal matrix \( S = PD^{-1}/2 \) to find,

\[
SLS^{-1} = \frac{1}{\epsilon} PD^{1/2} P^{-2}(D^{-1} K - I)P^{-1/2} = \frac{1}{\epsilon} P^{-1}(D^{-1/2} K D^{-1/2} - I)P^{-1} = \frac{1}{\epsilon} (S^{-1} KS^{-1} - P^{-2}).
\]

So we define the symmetric matrix \( \hat{L} = S^{-1} LS^{-1} \), and then note that setting \( U = \hat{L} \hat{U} \) we have,

\[
LU = S^{-1} \hat{L} S U = S^{-1} \hat{L} S S^{-1} \hat{U} = S^{-1} \hat{L} \hat{U} = S^{-1} \Lambda \hat{U} = \Lambda U,
\]
where \( \phi_i \) approximates an eigenfunction \( \phi \) evaluated at \( x_i, \phi(x_i) \), we can estimate the normalization factor as a Monte-Carlo integral given by,

\[
\|\phi\|_{L^2(q)} = \left( \int \phi(x)^2 q(x) \, dx \right)^{1/2} = \left( \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)^2 \right)^{1/2}.
\]

This implies that we should normalize the vector \( \vec{\phi} \) so that \( \|\vec{\phi}\|_2 = \left( \sum_{i=1}^{N} \phi_i^2 \right)^{1/2} = \sqrt{N} \). A further complication is the possibility of repeated eigenvalues, especially for a symmetric domain and potential. The numerical approximations to the eigenfunctions which correspond to a repeated eigenvalue can be any orthogonal transformation of the true eigenfunctions with the given eigenvalue. For visual comparison we compute this orthogonal transformation using the known eigenfunctions and apply it to the numerical eigenfunctions for each repeated eigenvalue.

The matrices formed above will all be \( N \times N \) where \( N \) is the number of data points. When \( N \) becomes large this quickly leads to large memory requirements. However, due to the exponential decay in the initial kernel \( K^S(x_i, x_j) \), we can replace these values by zero when \( x_j \) is far away from \( x_i \). In the examples below we will use the typical algorithm of taking the \( k \) nearest neighbors of each point \( x_i \) and allowing those to be the only nonzero values in the matrix. Notice that the kernel matrix formed using only the \( k \)-nearest neighbors may not be symmetric since the nearest neighbor relationship is not reflexive. If we are forming a kernel matrix \( K \) which should be symmetric using only the \( k \)-nearest neighbors, we always immediately replace \( K \) with \((K + K^\top)/2\) which is symmetric and still sparse. This sparse representation will only use \( O(Nk) \) memory rather than \( O(N^2) \) and will give similar results for \( \epsilon \) small enough that the truncated entries are already very close to zero. Since the initial kernel matrix is sparse, all the remaining matrices are simply multiplication and subtraction of sparse matrix by diagonal matrices and hence all the matrices constructed above will be sparse. We then use a sparse eigenvalue solver to find the desired number of eigenvectors of \( \tilde{L} \), with eigenvalues closest to \( \lambda_0 = 0 \).

5. Application to Ornstein-Uhlenbeck on \( \mathbb{R} \)

In this section we show the improvement in eigenvalue and eigenfunction estimation which is made possible by using variable bandwidth kernels rather than fixed bandwidth kernels on an unbounded manifold. We consider the backward Kolmogorov operator for the Ornstein-Uhlenbeck process on the real line. This process is driven by Brownian motion in a quadratic potential field \( U(x) = -\frac{1}{2}x^2 \) with invariant measure given by a standard normal distribution, \( q(x) \propto \exp(-U(x)) = \exp(-x^2/2) \).

In Section 2.1 we saw that the continuous theory of [4] easily extends to non-compact manifolds such as \( \mathbb{R} \) for functions \( f \) which are integrable with respect to the sampling measure \( q \). The problem arises when we try to
approximate the integral operators $G^\xi$ with discrete sums evaluated on a random data set. We first demonstrate that for a ‘nice’ sample set, the fixed bandwidth kernel can only approximate the operator $L_{\epsilon,\alpha,\beta}$ when $N$ is small and $\epsilon$ is carefully tuned. Rather than sampling randomly from a standard normal distribution, we first generate a ‘nice’ sample set with $N = 2000$ points by setting $\delta = (N + 1)^{-1}$ and generating a uniform grid $\{\hat{x}_i = \delta i\}_{i=1}^N$. We then apply the inverse of the cumulative distribution function to the uniform grid so that,

$$x_i = \sqrt{2} \text{erf}^{-1}(2\hat{x}_i - 1),$$

and $\{x_i\}$ have a standard normal distribution. In Figure 3, we compare the fourth eigenvector of the operator $L_{\epsilon,\alpha,\beta}^\xi$ with the analytic eigenfunction $H_3(x) = \frac{1}{\sqrt{6}}(x^3 - 3x)$ of the generator of the Ornstein-Uhlenbeck process. Notice that for a careful selection of $\epsilon$ the eigenvector approximation from the fixed bandwidth kernel ($\alpha = 1/2, \beta = 0$) agrees with the analytic eigenfunction. In contrast, the variable bandwidth kernel gives a much better approximation over a larger range of $\epsilon$ values. This demonstrates the first valuable aspect of variable bandwidth kernels, which is that they reduce the sensitivity to $\epsilon$. Moreover, when the true eigenfunction is unknown, it is difficult to determine which value of $\epsilon$ in the fixed bandwidth kernel is giving the best approximation. The fact that the variable bandwidth kernel gives a stable result, which persists across a large range of bandwidth choices, suggests that it may be possible to choose the bandwidth automatically based on stationarity of the solution.

In the limit of large data, the difficulty in applying the fixed bandwidth kernel becomes more severe. In Figure 4 we compare the variable bandwidth and fixed bandwidth approximations of the fourth eigenfunction for $N = 20000$ data points sampled according to (26). In this case, there is no value of $\epsilon$ which gives reasonable results for the fixed bandwidth kernel, whereas the variable bandwidth kernel once again returns excellent and stable results over a wide range of values of $\epsilon$. While it may seem counterintuitive that the fixed bandwidth kernel performs worse with more data, the error bound in Theorem 1 suggests exactly this effect. Since $c_2 > 0$ for the fixed bandwidth kernel, the third error bound diverges as the sampling density $q$ approaches zero. When the data set is small, it is unlikely that many of the samples are in areas of small sampling density. As the size of the data set is increased, the data set contains more points in areas of small sampling density and the minimum value of $q(x_i)$ decreases, causing the error bound to diverge.

One may hope to solve the issue of the error bound increasing in areas of small sampling by simply removing the outliers. However, this artificial modification of the data set implicitly creates an absorbing boundary condition at a virtual boundary between the remaining data and the removed outliers. This virtual boundary significantly affects the spectral properties of the discrete operator $L^\xi$ when $\beta = 0$. The structural error in the operator approximation due to the virtual boundary leads to extremely poor convergence properties for the eigenvalues of eigenfunctions of $L^\xi$ for...
$\beta = 0$. In Figure 5 we show that even with $N = 100000$ data points distributed according to (26), removing only 316 outlier points (those of smallest probability) has a significant effect on the eigenfunction approximation. This strategy may still have problems as $N$ increases, but even assuming that the trend in Figure 5 persists for large $N$, achieving the mean squared error of a variable bandwidth kernel applied to 1000 data points would require approximately $6 \times 10^7$ data points with a fixed bandwidth kernel using the outlier removal strategy.

We now consider the case of randomly sampled data, where the $x_i$ are independently sampled from a standard normal distribution, and demonstrate that fixed bandwidth kernels have even more significant limitations in this context. In Figure 6 we compare the fixed and variable bandwidth approximations with 10 randomly generated data sets of length $N = 20000$. Notice that none of the 10 fixed bandwidth approximations shown in Figure 6 converged to the correct eigenfunction. In contrast, the variable bandwidth kernel shows significant improvement with the increase in data in accordance with Theorem 1. Note that one of the data sets resulted in a particularly poor approximation even with the variable bandwidth kernel, but this does not contradict Theorem 1 since the error bounds are only obtained with high probability. In other words, for fixed $N$ a particularly bad random sample can lead to poor estimates even with the variable bandwidth.

6. Application to non-uniformly sampled unit circle

In this section we demonstrate that even on a compact manifold, when the sampling density $q$ is non-uniform, the variable bandwidth kernels still have many advantages. Consider a unit circle parameterized by $\theta \in [0, 2\pi]$ with the variable bandwidth $\Delta = \epsilon \frac{\sin(\theta)}{4N} (\cos(\theta) + \sin(\theta))$. We can produce a grid of $N = 1500$ points which have this distribution by numerically inverting the cumulative distribution function $F(\theta) = \frac{1}{2\pi} (2\theta + \sin(\theta))$ and applying it to a uniform grid $\{t_i = \delta\}_{i=1}^{2000}$ where $\delta = 1/1501$ so that $\theta_i = F^{-1}(t_i)$. We then produce a data set in $\mathbb{R}^2$ via the standard embedding $x_i = (\cos(\theta_i), \sin(\theta_i))^T$. Since the manifold is compact and the sampling density is bounded away from zero, we can approximate the Laplacian on the circle $\Delta = \frac{\partial^2}{\partial x^2}$ from the data set $\{x_i\}$ with either the fixed bandwidth kernel ($\alpha = 1/2, \beta = 0$) or the variable bandwidth kernel ($\alpha = 1/4, \beta = -1/2$). However, each algorithm requires tuning the nuisance parameter $\epsilon$, which can be very difficult for large data sets where the runtimes become restrictive. In Figure 7 we show that the variable bandwidth gives better results over a much larger range of $\epsilon$ values, which is an important consideration for practical applications. We then consider sensitivity to randomness in the data set by perturbing each $\theta_i$ by a uniform random variable from $[0, 0.5]$ and reduce modulo $2\pi$. The analysis is repeated for the randomized data set in Figure 7 and again the variable bandwidth kernel yields a better approximation with reduced sensitivity on the bandwidth $\epsilon$ near the optimal choice. Of course, the weakness of the variable bandwidth kernel is that it requires
knowledge of the intrinsic dimension $d$, which can be costly and difficult to estimate especially for noisy data. These results show that it may be worth the cost of estimating the dimension in order to recover a significantly improved approximation.

7. Conclusion

The theory developed above shows how to use variable bandwidth kernels to approximate the eigenfunctions of the Laplacian and the generators of gradient flow systems. We developed the general asymptotic expansion for the integral operators associated to the continuous variable bandwidth kernels of the form (4). This expansion reveals that a bandwidth function, $\rho$, changes the limiting operator to include a gradient term $(d+2)\nabla f \cdot \nabla \rho$. In the case of uniform sampling, this implies that the Laplacian Eigenmaps algorithm, with a kernel of the form (4) will no longer produce a Laplacian operator, but instead will produce the generator of a gradient flow with potential field $U(x) = -(d+2) \log(\rho)$.

As shown in (17), in the case of non-uniform sampling, we can remove the effect of the sampling, using $\alpha = 1$ as in Diffusion Maps [4]. This allows us to approximate the generator for any gradient flow system with known potential function $U$, by using the bandwidth function $\rho = \exp(-U/(d+2))$. Alternatively, by choosing the bandwidth function to be a power of the sampling density, we can recover a result similar to that of [4] but with a constant $c_1$ which depends on both the $\alpha$ normalization and the exponent $\beta$ used in the bandwidth function. Given a data set sampled from the invariant measure of a gradient flow system, this result allows us to use a large class of variable bandwidth kernels (define by the choice of $\beta$) to approximate the generator of the gradient flow system from which the data originates without any prior knowledge of the potential function.

While the theory developed in Section 2 describes the limiting operator for the continuous integral operator associated to a variable bandwidth kernel, in practice we are interested in approximating these integral operators from data. Often the data will be random samples of the density $q$, in which case we use discrete sums as Monte-Carlo approximations to the integral operators. By extending the analysis of Singer in [15], we showed in Section 3 that the bias error estimates may be unbounded as the sampling density $q$ approaches zero. Recall that in kernel density estimation problems, variable bandwidth kernels are known to give faster convergence rate and reduced sensitivity to the choice of bandwidth [16, 14], however it is still possible to use fixed bandwidth kernels to estimate densities. The theory developed in Section 3 and the numerical example demonstrated in Section 5 reveal that for operator approximation problems with densities that are not bounded away from zero, variable bandwidth kernels are not simply an improvement but are necessary for convergence. Fixed bandwidth kernels can be used for density estimation because the sampling density always appears in the numerator of the error bound, as shown in Section 3.1 whereas for operator approximation the density appears in the denominator of the error bound when $\beta = 0$ as shown in Section 3.2.
The main drawback of the variable bandwidth kernel algorithm is that it requires knowledge of the intrinsic dimension of the embedded manifold. This can be difficult in the presence of noise, and some techniques are given in [9, 11]. The dimension is required in two ways. First the equations for $c_1$ and $c_2$ depend on the dimension, and for $\beta \neq 0$ the dimension is required to find $c_1$ which determines the limiting operator. Second, the estimate $q^\alpha$ of the sampling density $q$, which is used to de-bias the operator using the $\alpha$ normalization of [4], requires a true variable bandwidth density estimate. Since the order zero term in the expansion of $G^d_\epsilon (q)$ is $mq \rho^d q$, we need to divide by $\rho^d$ to recover $q$, leading to the definition $q^\alpha (x_j) = \sum_i K^d_\epsilon (x_{ij}/\rho(x_j))^d$. We note that it is possible to perform the entire analysis without this division by $\rho^d$. We conducted a thorough analysis of this alternative normalization and found different formulas for $c_1$ and $c_2$. However in our analysis of this alternative formulation we found that the constraint $c_2 < 0$ would require $\frac{1}{\sigma} < \beta < 0$, which restricts $\beta$ significantly when $d$ is large, and means that once again the dimension must be known in order to choose $\beta$ in practice.

Acknowledgments

The research of J.H. is partially supported by the Office of Naval Research Grants N00014-11-1-0310, N00014-13-1-0797, MURI N00014-12-1-0912 and the National Science Foundation DMS-1317919. T. B. is supported under the ONR MURI grant N00014-12-1-0912.

References

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

[2] T. Berry, J. R. Cressman, Z. Gregurić Ferenček, and T. Sauer. Time-scale separation from diffusion-mapped delay coordinates. SIAM Journal of Applied Dynamical Systems, 12:618–649, 2013.

[3] R. Coifman, R. Erban, A. Singer, and I. Kevrekidis. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. PNAS, 106:113–127, 2009.

[4] R. Coifman and S. Lafon. Diffusion maps. Applied and Computational Harmonic Analysis, 21:5–30, 2006.

[5] R. Coifman, S. Lafon, M. Maggioni, B. Nadler, and I. Kevrekidis. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. SIAM Journal for Multiscale Modeling & Simulation, 7:842–864, 2008.

[6] R. Coifman, S. Lafon, B. Nadler, and I. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Applied and Computational Harmonic Analysis, 21:113–127, 2006.

[7] D. Giannakis and A. J. Majda. Time series reconstruction via machine learning: Revealing decadal variability and intermittency in the north pacific sector of a coupled climate model. In Conference on Intelligent Data Understanding (CIDU), pages 107–117, Mountain View, California, 2011.

[8] D. Giannakis and A. J. Majda. Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. Proceedings of the National Academy of Sciences, 109(7):2222–2227, 2012.
[9] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in \( \mathbb{R}^d \). In Proceedings of the 22nd International Conference on Machine Learning, ICML ’05, pages 289–296, New York, NY, USA, 2005. ACM.

[10] Matthias Hein, Jean-yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In Proceedings of the 18th Conference on Learning Theory (COLT), pages 470–485. Springer, 2005.

[11] A Little, Yoon-Mo Jung, and Mauro Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In Manifold learning and its applications : papers from the AAAI Fall Symposium, pages 26–33. AAAI, 2009.

[12] Emanuel Parzen. On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 33(3):1065–1076, 1962.

[13] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3):832–837, 1956.

[14] Stephan R. Sain and David W. Scott. On locally adaptive density estimation. Journal of the American Statistical Association, 91(436):1525–1534, 1996.

[15] A. Singer. From graph to manifold laplacian: The convergence rate. Appl. Comp. Harmonic Anal., 21:128–134, 2006.

[16] D. G. Terrell and D. W. Scott. Variable kernel density estimation. Annals of Statistics, 20:1236–1265, 1992.

[17] Daniel Ting, Ling Huang, and Michael I. Jordan. An analysis of the convergence of graph Laplacians. In Johannes Fürnkranz and Thorsten Joachims, editors, Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 1079–1086, Haifa, Israel, June 2010. Omnipress.