Available online at www.sciencedirect.com

SCIENCE DIRECT®

Appl. Comput. Harmon. Anal. 21 (2006) 5–30

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Diffusion maps

Ronald R. Coifman *, Stéphane Lafon [1]

Mathematics Department, Yale University, New Haven, CT 06520, USA

Received 29 October 2004; revised 19 March 2006; accepted 2 April 2006

Available online 19 June 2006

Communicated by the Editors

Abstract

In this paper, we provide a framework based upon diffusion processes for finding meaningful geometric descriptions of data sets. We show that eigenfunctions of Markov matrices can be used to construct coordinates called *diffusion maps* that generate efficient representations of complex geometric structures. The associated family of *diffusion distances*, obtained by iterating the Markov matrix, defines multiscale geometries that prove to be useful in the context of data parametrization and dimensionality reduction. The proposed framework relates the spectral properties of Markov processes to their geometric counterparts and it unifies ideas arising in a variety of contexts such as machine learning, spectral graph theory and eigenmap methods.

Keywords: Diffusion processes; Diffusion metric; Manifold learning; Dimensionality reduction; Eigenmaps; Graph Laplacian

1. Introduction

Dimensionality reduction occupies a central position in many fields such as information theory, where it is related to compression and coding, statistics, with latent variables, as well as machine learning and sampling theory. In essence, the goal is to change the representation of data sets, originally in a form involving a large number of variables, into a low-dimensional description using only a small number of free parameters. The new representation should describe the data in a faithful manner, by, say, preserving some quantities of interest such as local mutual distances. Analogous to the problem of dimensionality reduction is that of finding meaningful structures in data sets. The idea here is slightly different and the goal is to extract relevant features out of the data in order to gain insight and understanding of the phenomenon that generated the data.

In order to achieve any of these two goals, numerous data mining and machine learning techniques rely on graph-based algorithms. In terms of data structures, graphs offer an advantageous compromise between their simplicity, interpretability and their ability to represent complex relationships between data points. Weighted graphs are usually

* Corresponding author.
E-mail addresses: coifman@math.yale.edu (R.R. Coifman), stephane.lafon@gmail.com (S. Lafon).
[1] Now with Google Inc.

1063-5203/$ – see front matter
doi:10.1016/j.acha.2006.04.006

employed to represent a notion of geometry[2] based on the local similarity or interaction between the data points. In many situations, each data sample is represented by a collection of numerical attributes, and in this case, the condition for two nodes to be connected (and the strength of such a connection) is based on the proximity of the corresponding data points in the feature space. In the context of networks (e.g., social, computer, communication or transportation networks), the data naturally lend themselves to graph modeling. For instance, graph methods play a major role in the quantitative modeling of social networks [1]. When combined with Markov chain techniques, graph-based methods can be extremely successful. In particular, in the arena of classification and clustering, random walks on graphs have proven to be very efficient at finding relevant structures in complex geometries. For instance, in [2], the L^1 distance between probabilities of transition is used as a metric between data points, and this metric is then employed to induce class labels. This technique is shown to be quite successful when classes have nonlinear shapes. In the field of spectral clustering, a Markov chain is constructed over the graph of the data, and the sign of the values of the top nonconstant eigenvector of the corresponding transition matrix is used for finding clusters and computing cuts [3,4].

If one uses the values of this eigenvector (and not only their sign), one obtains a score for each data point. This leads to all kinds of ranking techniques. In particular, the celebrated PageRank algorithm [5,6] essentially uses the stationary distribution of a random walk on the link structure of the web in order to rank web documents in terms of relative importance. Several variations around this theme appear in the literature [7,8], where, just like PageRank, various techniques provide a ranking function based on the top eigenvector of a Markov matrix [9,10].

This type of approach was later generalized to using higher-order eigenvectors. The field of applications then goes beyond the ideas of clustering and ranking, as using multiple eigenvectors allows to speak of *parametrization* of data sets. A great deal of attention has been recently paid to the so-called "kernel eigenmap methods" such as local linear embedding [21], Laplacian eigenmaps [11], Hessian eigenmaps [18] and local tangent space alignment [25]. The remarkable idea emerging from these papers is that eigenvectors of Markov matrices can be thought of as coordinates on the data set. Therefore, the data, originally modeled as a graph, can be represented (embedded) as a cloud of points in a Euclidean space. In addition, the hope is that this new representation will capture the main structures of the data in a few dimensions, hence achieving dimensionality reduction. These algorithms exhibit two major advantages over classical dimensionality reduction methods (such as principal component analysis or classical multidimensional scaling): they are nonlinear, and they preserve local structures. The first aspect is essential as most of the time, in their original form, the data points do not lie on linear manifolds. The second point is the expression of the fact that in many applications, distances of points that are far apart are meaningless, and therefore need not be preserved.

In this paper, we show that all these kernel eigenmap methods constitute special cases of a general framework based on diffusion processes. We use the eigenfunctions of a Markov matrix defining a random walk on the data to obtain new descriptions of data sets (subset of $\mathbb{R}^n$, graphs) via a family of mappings that we term "diffusion maps." These mappings embed the data points into a Euclidean space in which the usual distance describes the relationship between pairs of points in terms of their connectivity. This defines a useful distance between points in the data set that we term "diffusion distance." The approach that we present generalizes the classical Newtonian paradigm in which local infinitesimal rules of transition of a system lead to global macroscopic descriptions by integration. We obtain different geometric representations of the data set by iterating the Markov matrix of transition, or equivalently, by running the random walk forward, and the diffusion maps are precisely the tools that allow us to relate the spectral properties of the diffusion process to the geometry of the data set. In particular, we do not obtain one representation of the geometry for the set, but a multiscale family of geometric representations corresponding to descriptions at different scales. This paper is organized as follows: in Section 2, we introduce the general framework by defining the diffusion maps and diffusion distances, and we show the relation with other kernel methods. In Section 3, we focus on the specific example of subsets of the Euclidean space $\mathbb{R}^n$. In particular, we construct a one-parameter family of diffusions that proves useful in the study of some stochastic dynamical systems, as well as for recovering the geometry of the data via the computation of the Laplace–Beltrami operator on manifolds. In Section 4, we extend the idea of anisotropic diffusion by showing that the construction of diffusion kernels can be data-driven and task-specific. Section 5 addresses the practical issue of dealing with finite data, as well as the robustness to noise. Section 6 explores various manners to combine several scales and describes other types of diffusion distances.

[2] In this article, the word "geometry" refers to a set of rules describing the relationships between data points. For instance, "being close to a data point x" is such a rule.

2. Diffusion maps

2.1. Construction of a random walk on the data

Let $(X, \mathcal{A}, \mu)$ be a measure space. The set X is the data set and μ represents the distribution of the points on X. In addition to this structure, suppose that we are given a "kernel" $k : X \times X \to \mathbb{R}$ that satisfies:

- k is symmetric: $k(x, y) = k(y, x)$,
- k is positivity preserving: $k(x, y) \geqslant 0$.

This kernel represents some notion of affinity or similarity between points of X as it describes the relationship between pairs of points in this set and in this sense, one can think of the data points as being the nodes of a symmetric graph whose weight function is specified by k. The kernel constitutes our prior definition of the *local* geometry of X, and since a given kernel will capture a specific feature of the data set, its choice should be guided by the application that one has in mind. This is a major difference with global methods like principal component analysis or multidimensional scaling where all correlations between data points are taken into account. Here, we start from the idea that, in many applications, high correlation values constitute the only meaningful information on the data set. Later in this paper, we illustrate this point by defining a one-parameter family of kernels, and we show that the corresponding diffusions can be used to analyze the geometry, the statistics or some dynamics of the data.

The reader might notice that the conditions on k are somewhat reminiscent of the definition of symmetric diffusion semi-groups [22]. In fact, to any reversible Markov process, one can associate a symmetric graph, and as we now explain, the converse is also true: from the graph defined by (X, k), one can construct a reversible Markov chain on X. The technique is classical in various fields, and is known as the normalized graph Laplacian construction [3]:

$$\text{set} \quad d(x) = \int_X k(x, y)\, \mathrm{d}\mu(y)$$

to be a local measure of the volume (or degree in a graph) and define

$$p(x, y) = \frac{k(x, y)}{d(x)}.$$

Although the new kernel p inherits the positivity-preserving property, it is no longer symmetric. However, we have gained a conservation property:

$$\int_X p(x, y)\, \mathrm{d}\mu(y) = 1.$$

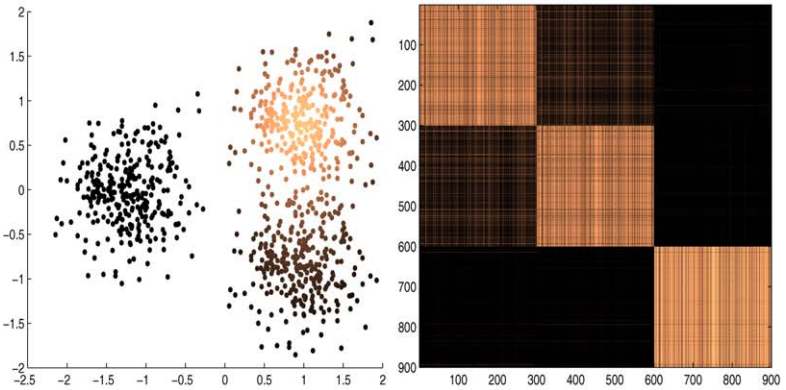
This means that p can be viewed as the transition kernel of a Markov chain on X, or, equivalently, the operator P defined by

$$Pf(x) = \int_X a(x, y) f(y)\, \mathrm{d}\mu(y)$$

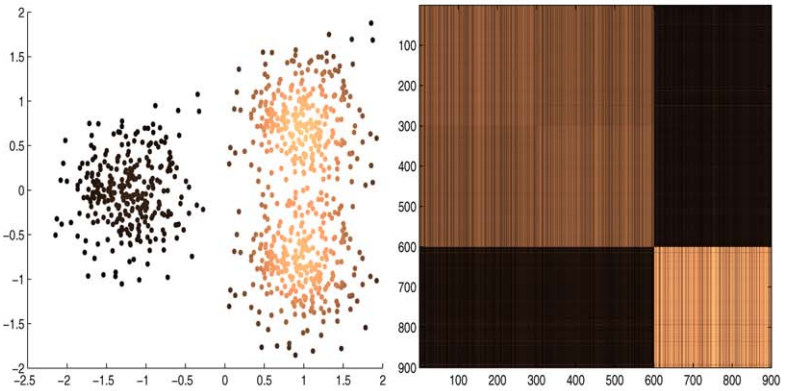
preserves constant functions (it is an averaging or diffusion operator).

2.2. Powers of P and multiscale geometric analysis of X

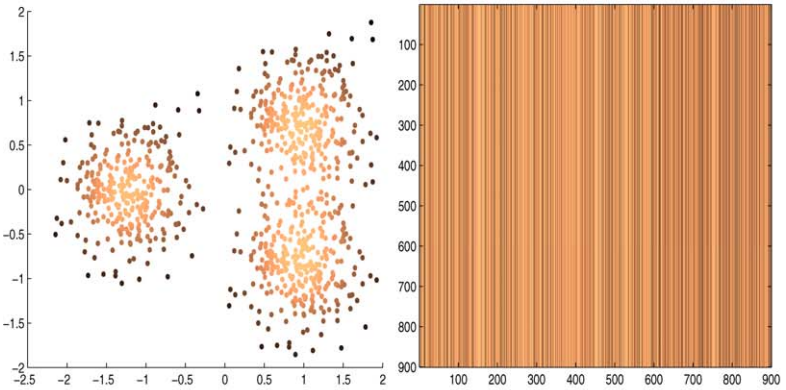
From a data analysis point of view, the reason for studying this Markov chain is that the matrix P contains geometric information about the data set X. Indeed, the transitions that it defines directly reflect the local geometry defined by the immediate neighbors of each node in the graph of the data. In other words, $p(x, y)$ represents the probability of transition in one time step from node x to node y and it is proportional to the edge-weight $k(x, y)$. For $t \geqslant 0$, the probability of transition from x to y in t time steps is given by $p_t(x, y)$, the kernel of the tth power P^t of P. One of the main ideas of the diffusion framework is that running the chain forward in time, or equivalently, taking larger



(a) $t = 8$



(b) $t = 64$



(c) $t = 1024$

Fig. 1. Diffusion at times $t = 8$, $t = 64$ and $t = 1024$ over a set containing 3 clusters. The left column represents the set, and the color encodes the intensity of diffusion from a fixed given point, that is, it corresponds to a given row of the corresponding power of P. The right column is a plot of the transition matrices P^8, P^{64} and P^{1024}. Points in X are ordered so that the first 300 roughly correspond to the first cluster, the next 300 are in the second cluster and so on.

powers of P, will allow us to integrate the local geometry and therefore will reveal relevant geometric structures of X at different scales.

An illustration of this idea is provided in Fig. 1. We generated a set X of 900 points in the plane. This set is formed by the union of 3 clusters. From this set, we built a graph with Gaussian weights $e^{-\|x_i - x_j\|^2/\varepsilon}$ with $\varepsilon = 0.7$, and formed the corresponding Markov matrix P. On this figure we plotted several powers of the matrix P, namely at times $t = 8$,

$t = 64$ and $t = 1024$. The block structure[3] of these powers clearly reveals the multiscale structure of the data: at $t = 8$, the set appears to be made of 3 distinct clusters. At $t = 64$, the two closest clusters have merged, and the data set is made of 2 clusters. Last, at $t = 1024$, all clusters have merged. Note also that P^{1024} appears to be (numerically) of rank one, as we have the approximate equality $p_{1024}(x, y) \simeq \pi(y)$ for all x and y. The key idea in this example is that the very notion of a cluster from a random walk point of view is *a region in which the probability of escaping this region is low*. This simple illustration also emphasizes the fact that, in addition to being the time parameter, t plays the role of a scale parameter.

2.3. Spectral analysis of the Markov chain

We conclude from the previous section that powers of P constitute an object of interest for the study of the geometric structures of X at various scales. A classical way to describe the powers of an operator is to employ the language of spectral theory, namely eigenvectors and eigenvalues. Although for general transition matrices of Markov chains, the existence of a spectral theory is not guaranteed, the random walk that we have constructed exhibits very particular mathematical properties:

- The Markov chain has a stationary distribution given by
$$\pi(y) = \frac{\mathrm{d}(y)}{\sum_{z\in X} \mathrm{d}(z)}.$$
If the graph is connected, which we now assume, then the stationary distribution is unique.
- The chain is reversible, i.e., it follows the detailed balance condition:
$$\pi(x)p(x, y) = \pi(y)p(y, x). \tag{1}$$
- If X is finite and the graph of the data is connected, then the chain is ergodic.[4]

Equation (1) plays a central role as it opens the door to a spectral analysis of the Markov chain. Under mild additional assumptions on k described in Appendix A, P has a discrete sequence of eigenvalues $\{\lambda_l\}_{l\geqslant 0}$ and eigenfunctions $\{\psi_l\}_{l\geqslant 0}$ such that $1 = \lambda_0 > |\lambda_1| \geqslant |\lambda_2| \geqslant \cdots$ and

$$P\psi_l = \lambda_l \psi_l.$$

2.4. Diffusion distances and diffusion maps

In this paragraph, we relate the spectral properties of the Markov chain to the geometry of the data set X. As previously mentioned, the idea of defining a random walk on the data set relies on the following principle: the kernel k specifies the local geometry of the data and captures some geometric feature of interest. The Markov chain defines fast and slow directions of propagation, based on the values taken by the kernel, and as one runs the walk forward, the local geometry information is being propagated and accumulated the same way local transitions of a system (given by a differential equation) can be integrated in order to obtain a global characterization of this system.

Running the chain forward is equivalent to computing the powers of the operator P. For this computation, we could, in theory, use the eigenvectors and eigenvalues of P. Instead, we are going to directly employ these objects in order to characterize the geometry of the data set X.

We start by introducing the family of *diffusion distances* $\{D_t\}_{t\in\mathbb{N}}$ given by

$$D_t(x, y)^2 \triangleq \left\| p_t(x, \cdot) - p_t(y, \cdot) \right\|^2_{L^2(X, \mathrm{d}\mu/\pi)} = \int_X \left(p_t(x, u) - p_t(y, u) \right)^2 \frac{\mathrm{d}\mu(u)}{\pi(u)}.$$

[3] It might seem that the block structure depends on the specific ordering of the points. However, as we show later, this issue is overcome by the introduction of the diffusion coordinates. These coordinates automatically organize the data regardless of the ordering.

[4] The state space of this Markov chain being finite, the ergodicity follows from the irreducibility and aperiodicity of the random walk. The irreducibility results from the graph being connected. In addition, since $k(x, x)$ represents the affinity of x with itself, one can reasonably assume that $k(x, y) > 0$, which implies that $p(x, x) > 0$, from which the aperiodicity follows.

In other words, $D_t(x, y)$ is a functional weighted L^2 distance between the two posterior distributions $u \mapsto p_t(x, u)$ and $u \mapsto p_t(y, u)$. For a fixed value of t, D_t defines a distance on the set X. By definition, the notion of proximity that it defines reflects the connectivity in the graph of the data. Indeed, $D_t(x, y)$ will be small if there is a large number of short paths connecting x and y, that is, if there is a large probability of transition from x to y and vice versa. In addition, as previously noted, t plays the role of a scale parameter. Therefore we underline three main interesting features of the diffusion distance:

- Since it reflects the connectivity of the data at a given scale, points are closer if they are highly connected in the graph. Therefore, this distance emphasizes the notion of a cluster.
- The quantity $D_t(x, y)$ involves summing over all paths of length t connecting x to y and y to x. As a consequence, this number is very robust to noise perturbation, unlike the geodesic distance.
- From a machine learning point of view, the same observation allows us to conclude that this distance is appropriate for the design of inference algorithms based on the majority of preponderance: this distance takes into account all evidences relating x and y.

As shown in Appendix A, $D_t(x, y)$ can be computed using the eigenvectors and eigenvalues of P:

$$D_t(x, y) = \left(\sum_{l \geqslant 1} \lambda_l^{2t} \big(\psi_l(x) - \psi_l(y)\big)^2 \right)^{\frac{1}{2}}.$$

Note that as ψ_0 is constant, we have omitted the term corresponding to $l = 0$.

Now, as previously mentioned, the eigenvalues $\lambda_1, \lambda_2, \ldots$, tend to 0 and have a modulus strictly less than 1. As a consequence, the above sum can be computed to a preset accuracy $\delta > 0$ with a finite number of terms: if we define

$$s(\delta, t) = \max\big\{ l \in \mathbb{N} \text{ such that } |\lambda_l|^t > \delta |\lambda_1|^t \big\},$$

then, up to relative precision δ, we have

$$D_t(x, y) = \left(\sum_{l=1}^{s(\delta,t)} \lambda_l^{2t} \big(\psi_l(x) - \psi_l(y)\big)^2 \right)^{\frac{1}{2}}.$$

We therefore introduce the family of *diffusion maps* $\{\Psi_t\}_{t \in \mathbb{N}}$ given by

$$\Psi_t(x) \triangleq \begin{pmatrix} \lambda_1^t \psi_1(x) \\ \lambda_2^t \psi_2(x) \\ \vdots \\ \lambda_{s(\delta,t)}^t \psi_{s(\delta,t)}(x) \end{pmatrix}.$$

Each component of $\Psi_t(x)$ is termed *diffusion coordinate*. The map $\Psi_t : X \to \mathbb{R}^{s(\delta,t)}$ embeds the data set into a Euclidean space of $s(\delta, t)$ dimensions.

The connection between diffusion maps and diffusion distances can be summarized as follows:

Proposition 1. *The diffusion map Ψ_t embeds the data into the Euclidean space $\mathbb{R}^{s(\delta,t)}$ so that in this space, the Euclidean distance is equal to the diffusion distance* (*up to relative accuracy δ*)*, or equivalently,*

$$\big\| \Psi_t(x) - \Psi_t(y) \big\| = D_t(x, y).$$

Note that using the full eigenvector expansion in the sum above proves that the diffusion distance D_t is a metric distance on X.

2.5. Parametrization of data and dimensionality reduction

The previous proposition states that the diffusion maps offer a representation of the data as a cloud of points in a Euclidean space. This representation is characterized by the fact the distance between two points is equal to the diffusion distance in the original description of the data. Therefore, the mapping Ψ_t reorganizes the data points according to their mutual diffusion distances.

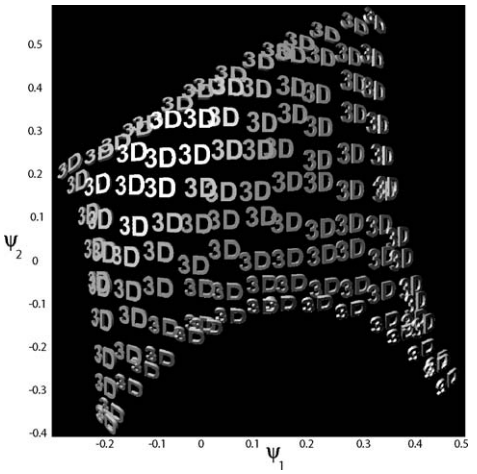


Fig. 2. The set of images reorganized by the two eigenvectors ψ_1 and ψ_2. We recover the natural organization dictated by the angles of rotation.

An illustration of the organizational power of the diffusion maps is shown in Fig. 2. We generated a collection of images of the word "3D" viewed under different angles and given in no particular order. We formed a diffusion matrix P based on a Gaussian-weighted graph and computed the diffusion coordinates. The figure shows the plot of the data in the first two eigenfunctions (ψ_1, ψ_2). The result demonstrates the organizational capability of these coordinates as they recover the natural parameter that generated the data, namely the two angles of variation. This automatic organization of the points is useful to learn nonlinear global parameters governing the geometry of X.

In general, for a given time t, the number of eigenvectors used for parametrizing the data is equal to the number of eigenvalues to the powers of t that have a magnitude greater than a given threshold δ. Therefore, the dimensionality of the embedding depends on both t and the decay of the spectrum of P. One extreme case corresponds to a graph where all the nodes are disconnected. This leads to P being equal to the identity operator and thus to a flat spectrum. At the other end of the family of graphs, consider a graph where all nodes are connected all the other nodes with weights equal to 1. In this case, P has one eigenvalue equal to 1, and all other eigenvalues are equal to 0 (we obtain the fastest decay possible for a diffusion operator). The decay of the spectrum is therefore a measure of the connectivity of points in the graph. In addition, many graphs formed from real-life data sets lie in between these two extreme cases. For instance, in the case when the data approximately lie on a submanifold, then, as we show later in Section 3, P is used as an approximation to the heat kernel on the submanifold. As we know from asymptotic expansion of the trace of this operator [30], the spectrum of the heat kernel decays smoothly (see Fig. 3 for a typical example of a graph of the spectrum of P on a submanifold), and the rate of decay depends on the intrinsic dimension of the submanifold as well as other quantities such as its volume, the area of its boundary, and other topological quantities such as the characteristic of the submanifold.

As a consequence, the diffusion maps allow to achieve dimensionality reduction, and the dimension of the embedding depends on both the geometry and the topology of the data set. In particular, if X is a discretized submanifold, the dimension of the embedding can be different from that of the submanifold.

2.6. Compression of operators and Heisenberg principle

The principle of studying the spectral properties of a Markov chain defined on a set X actually combines ideas from potential theory, spectral geometry and the study of partial differential operators. From potential theory, we know that the singularities of a domain (cusps, corners) are reflected in the behavior of the solution of Dirichlet and Neumann problems. Spectral geometry asks the question of whether the geometry of a Riemannian manifold is determined by the spectrum of Laplace operator [13]. More generally, the study of spectral asymptotics for partial differential operators relates geometric characteristics of a domain X to the growth of the eigenvalues of such operators. The common denominator of these ideas is that the geometry of a set X can be studied through the analysis of spaces of functions defined on X and linear operators over those spaces. The spectral analysis of the diffusion operator P serves as a tool for the geometric analysis of X.

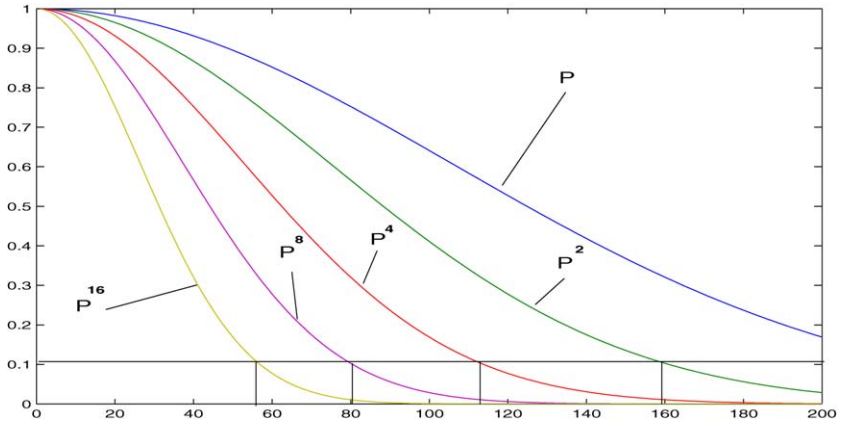


Fig. 3. The numerical rank of P^t decays as t increases, and so does the number of significant eigenvalues.

The rows of P^t give the probability of transition in t steps of the Markov chain, but they also have a dual interpretation as functions defined on the data set X. More precisely, if $x \in X$ is fixed, then $p_t(x, \cdot)$ is a bump function centered at x and of width increasing with t. The diffusion distance, in addition to being a distance between points is also a distance between these bumps:

$$D_t^2(x, y) = \int_X \left| p_t(x, u) - p_t(y, u) \right|^2 \frac{\mathrm{d}\mu(u)}{\pi(u)}.$$

When t gets larger, the size of the support of $p_t(x, \cdot)$ increases, and as already noted, the number of eigenfunctions necessary to synthesize this bump gets smaller. This number is related to the minimum number of bumps necessary to cover to the set X (like in Weyl's asymptotic law for the decay of the spectrum, see [14,15]). As a consequence, the set of these bumps can be downsampled (see [17] for an efficient implementation of this downsampling) in order to compress the operator P^t and the linear span of this downsampled set is therefore a space of functions with low frequencies. By a change of basis, this implies that the eigenfunctions corresponding to eigenvalues at the beginning of the spectrum (close to $\lambda_0 = 1$) have a low-frequency content, and as one goes further down in the spectrum, the eigenfunctions become increasingly oscillatory. This relation between the size of the support of the bumps, the position of the corresponding eigenvalues in the spectrum and their frequency content is the expression of some form of the Heisenberg principle induced by the diffusion operator P.

On the one hand, this simple observation has powerful implications in numerical analysis. Indeed, the action of operator P and its powers therefore exhibit a multiscale structure of which one can take advantage in order to rapidly compute functions of this operator [16,17]. On the other hand, the family of diffusion distances and maps form a flexible tool for data analysis as the different geometric information at all scales can be combined (see Section 6).

2.7. Link with other kernel eigenmap methods

As mentioned earlier, the so-called kernel methods from machine learning can be analyzed in the context of diffusion maps. In [11], Belkin shows that the LLE algorithm is equivalent to finding the eigenfunctions of the square of the normalized graph Laplacian, which is the same as diagonalizing the graph Laplacian itself. In [20], it is shown that all kernel-based manifold learning methods are special cases of kernel PCA. We now show how they are related to diffusion maps.

In one way or another, kernel eigenmap methods all aim at solving the following problem:

$$\min_{Q_2(f)=1} Q_1(f), \quad \text{where } Q_1(f) = \sum_{x \in X} Q_x(f),$$

where Q_2 and $\{Q_x,\ x \in X\}$ are symmetric positive semi-definite quadratic forms, acting on functions f defined on X. The quadratic forms Q_x are local in that they measure local variations of f; in particular Q_1 vanishes on constant functions and its matrix is sparse. Typically, $Q_x(f)$ is the square of the gradient of f at x, or the squared Fröbenius

norm of the Hessian of f at x. The quadratic form Q_2 acts as a normalization for f and in general, its matrix is diagonal.

For example, with the construction of the random walk from a kernel k (as explained in the previous paragraphs),

$$Q_1(f) = \sum_{x\in X}\sum_{y\in X} k(x,y)\big(f(x)-f(y)\big)^2$$

and

$$Q_2(f) = \sum_{x\in X} v(x) f(x)^2.$$

Generally, the problem is solved by computing the solution of $Q_1 f = \lambda Q_2 f$ (note that this system is sparse), which is equivalent to solving

$$Q_2^{-1} Q_1 f = \lambda f$$

and these eigenfunctions are then used to embed the data points. The operator $Q_2^{-1}Q_1$ generally represents the discretization of a differential operator of even order and assuming that the spectrum of $Q_2^{-1}Q_1$ is between 0 and 1, it is the infinitesimal generator of the diffusion defined by $e^{-tQ_2^{-1}Q_1}$.

3. Anisotropic diffusions for points in $\mathbb{R}^n$

We now focus on the case of data points in the Euclidean space $\mathbb{R}^n$. Examples include data sets approximating Riemannian submanifolds as well as data points sampled from the equilibrium distribution of stochastic dynamical systems. Manifold models are important in many applications, such as image analysis and computer vision [27,28], and one is generally interested in obtaining a low-dimensional representation of the data set, such as a coordinate system. Now, since the sampling of the data is generally not related to the geometry of the manifold, one would like to recover the manifold structure regardless of the distribution of the data points. In the case when the data points are sampled from the equilibrium distribution of a stochastic dynamical system, the situation is quite different as the density of the points is a quantity of interest, and therefore, cannot be gotten rid of. Indeed, for some dynamical physical systems, regions of high density correspond to minima of the free energy of the system. Consequently, the long-time behavior of the dynamics of this system results in a subtle interaction between the statistics (density) and the geometry of the data set.

It is very tempting to process data sets in $\mathbb{R}^n$ by considering the graph formed by the data points and whose weights are given by some isotropic kernel, e.g., $k_\varepsilon(x,y) = e^{-\|x-y\|^2/\varepsilon}$ for some carefully chosen scale parameter ε. In [11], Belkin and Niyogi suggest to compute the normalized graph Laplacian from this kernel and use the spectral properties of the corresponding diffusion to cluster and organize the data. Although the virtues of this type of approach are well known for a general graph (see [4,26]), more can be said for the special case of points in the Euclidean space. In particular, what is the influence of the density of the points and of the geometry of the possible underlying data set over the eigenfunctions and spectrum of the diffusion?

To address this type of question, we now introduce a family of anisotropic diffusion processes that are all obtained as small-scale limits of a graph Laplacian jump process. This family is parameterized by a number $\alpha \in \mathbb{R}$ which can be tuned up to specify the amount of influence of the density in the infinitesimal transitions of the diffusion. The crucial point is that the graph Laplacian normalization is *not* applied on a graph with isotropic weights, but rather on a renormalized graph. Three values of the parameter α are particularly interesting:

- When $\alpha = 0$, the diffusion reduces to that of the classical normalized graph Laplacian normalization applied to the graph with isotropic weights, e.g., $e^{-\|x_i-x_j\|^2/\varepsilon}$. The influence of the density is maximal in this case.
- For the intermediate case $\alpha = \frac{1}{2}$, the Markov chain is an approximation of the diffusion of a Fokker–Planck equation, allowing to approximate the long-time behavior or the point distribution of a system described by a certain stochastic differential equation.
- When $\alpha = 1$, and if the points approximately lie on a submanifold of $\mathbb{R}^n$, one obtains an approximation of the Laplace–Beltrami operator. In this case, one is able to recover the Riemannian geometry of the data set, regardless of the distribution of the points. This case is particularly important in many applications.

In the following, we start by explaining the construction of this family of diffusions and then we study each of the above special cases separately. Let us fix the notation and review some notions related to the heat propagation on submanifolds. Let $\mathcal{M}$ be a compact C^∞ submanifold of $\mathbb{R}^n$. The heat diffusion on $\mathcal{M}$ is the diffusion process whose infinitesimal generator is the Laplace–Beltrami operator Δ (we adopt the convention that this operator is positive semi-definite). Let the Neumann heat kernel be denoted $e^{-t\Delta}$. The operator Δ has eigenvalues and eigenfunctions on $\mathcal{M}$:

$$\Delta\phi_l = \nu_l^2\phi_l,$$

where ϕ_l verifies the Neumann condition $\partial\phi_l = 0$ at the boundary ∂M. These eigenfunctions form a Hilbert basis of $L^2(\mathcal{M}, dx)$. Let

$$E_K = \mathrm{Span}\{\phi_l,\ 0 \leqslant l \leqslant K\}$$

be the linear span of the first $K+1$ Neumann eigenfunctions. Another expression for the Neumann heat kernel is given by

$$e^{-t\Delta} = \lim_{s\to+\infty}\left(I - \frac{\Delta}{s}\right)^{st} = \sum_{l\geqslant 0} e^{-t\nu_l^2}\phi_l(x)\phi_l(y).$$

We will assume that the data set X is the entire manifold (as later in this paper we address the question of finite sets approximating $\mathcal{M}$). Let $q(x)$ be the density of the points on $\mathcal{M}$.

3.1. Construction of a family of diffusions

There are two steps in the algorithm: one first renormalizes the rotation-invariant weight into an anisotropic kernel, and then one computes the normalized graph Laplacian diffusion from this new graph.

Construction of the family of diffusions

(1) Fix $\alpha \in \mathbb{R}$ and a rotation-invariant kernel $k_\varepsilon(x, y) = h\left(\frac{\|x-y\|^2}{\varepsilon}\right)$.
(2) Let

$$q_\varepsilon(x) = \int_X k_\varepsilon(x, y)q(y)\,\mathrm{d}y$$

and form the new kernel

$$k_\varepsilon^{(\alpha)}(x, y) = \frac{k_\varepsilon(x, y)}{q_\varepsilon^\alpha(x)q_\varepsilon^\alpha(y)}.$$

(3) Apply the weighted graph Laplacian normalization to this kernel by setting

$$d_\varepsilon^{(\alpha)}(x) = \int_X k_\varepsilon^{(\alpha)}(x, y)q(y)\,\mathrm{d}y$$

and by defining the anisotropic transition kernel

$$p_{\varepsilon,\alpha}(x, y) = \frac{k_\varepsilon^{(\alpha)}(x, y)}{d_\varepsilon^{(\alpha)}(x)}.$$

Note that, up to a multiplicative factor, the quantity $q_\varepsilon(x)$ is an approximation of the true density $q(x)$. Let $P_{\varepsilon,\alpha}$ be defined by

$$P_{\varepsilon,\alpha}f(x) = \int_X p_{\varepsilon,\alpha}(x, y)f(y)q(y)\,\mathrm{d}y.$$

Our main result[5] concerns the infinitesimal generator of the corresponding diffusion as $\varepsilon \to 0$:

Theorem 2. *Let*

$$L_{\varepsilon,\alpha} = \frac{I - P_{\varepsilon,\alpha}}{\varepsilon}$$

be the infinitesimal generator of the Markov chain. Then for a fixed $K > 0$, we have on E_K

$$\lim_{\varepsilon\to 0} L_{\varepsilon,\alpha} f = \frac{\Delta(f q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}} f.$$

In other words, the eigenfunctions of $P_{\varepsilon,\alpha}$ can be used to approximate those of the following symmetric Schrödinger operator:

$$\Delta\phi - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\phi,$$

where $\phi = f q^{1-\alpha}$.

The proof is given in Appendix B.

3.2. The case $\alpha = 0$: normalized graph Laplacian on isotropic weights

Setting $\alpha = 0$ comes down to computing the normalized graph Laplacian on a graph with isotropic (e.g., Gaussian) weights. From the previous theorem, the corresponding infinitesimal operator is given by

$$\Delta\phi - \frac{\Delta q}{q}\phi.$$

Note that when the density q is uniform on $\mathcal{M}$, then the potential term vanishes.[6] This is consistent with Belkin's result [11] saying that in this situation, the graph Laplacian normalization yields an approximation of the Laplace–Beltrami operator on $\mathcal{M}$. However our analysis also reveals that in the general case, this no longer holds, and simple calculations show that the influence of the density term can be quite important.

3.3. The case $\alpha = \frac{1}{2}$: Fokker–Planck diffusion

When $\alpha = \frac{1}{2}$, the asymptotic infinitesimal generator reduces to

$$\Delta\phi - \frac{\Delta(\sqrt{q})}{\sqrt{q}}\phi.$$

Let us write $q = e^{-U}$, then the generator becomes

$$\Delta\phi - \left(\frac{\|\nabla U\|^2}{4} - \frac{\Delta U}{2}\right)\phi.$$

It is shown in [29] that a simple conjugation of this specific Schrödinger operator leads to the forward Fokker–Plank equation

$$\frac{\partial q}{\partial t} = \nabla\cdot(\nabla q + q\nabla U),$$

where $q(x,t)$ represents the density of points at position x and time t of a dynamical system satisfying the Langevin equation

[5] Without any loss of generality, we assume that function h has a zeroth moment equal to 1 and second moment equal to 2 (see the proof of the theorem on p. 27).

[6] In practice, a perfectly uniform density is difficult to achieve, if possible at all. Therefore this scheme cannot really be used to approximate Δ as the slightest nonconstant mode in q is amplified in the factor $\Delta q/q$.

$$\dot{x} = -\nabla U(x) + \sqrt{2}\dot{w}, \tag{2}$$

where w is an n-dimensional Brownian motion.

The implication of this observation is that this normalization can be used for the analysis of stochastic dynamical systems governed by the Langevin equation (2). In particular, it shown in [29] that this normalization is a powerful tool for the study of systems exhibiting different time scales: a short-time scale corresponding to the Brownian fluctuations, and a long-time scale corresponding to the drift induced by the vector field ∇U.

3.4. The case $\alpha = 1$: approximation of the heat kernel

Finally, when $\alpha = 0$, we obtain the following important result:

Proposition 3. *We have*:

$$\lim_{\varepsilon \to 0} L_{\varepsilon,1} = \Delta.$$

Furthermore, for any $t > 0$, the Neumann heat kernel $e^{-t\Delta}$ can be approximated on $L^2(\mathcal{M})$ by $P_{\varepsilon,1}^{\frac{t}{\varepsilon}}$:

$$\lim_{\varepsilon \to 0} P_{\varepsilon,1}^{\frac{t}{\varepsilon}} = e^{-t\Delta}.$$

The proof is given in Appendix B.

By setting $\alpha = 1$, the infinitesimal generator is simply the Laplace–Beltrami operator Δ. In other words, the Markov chain converges to the Brownian motion on $\mathcal{M}$. As a consequence, this normalization removes all influence of the density and recovers the Riemannian geometry of the data set. This procedure therefore allows to separate the dis-

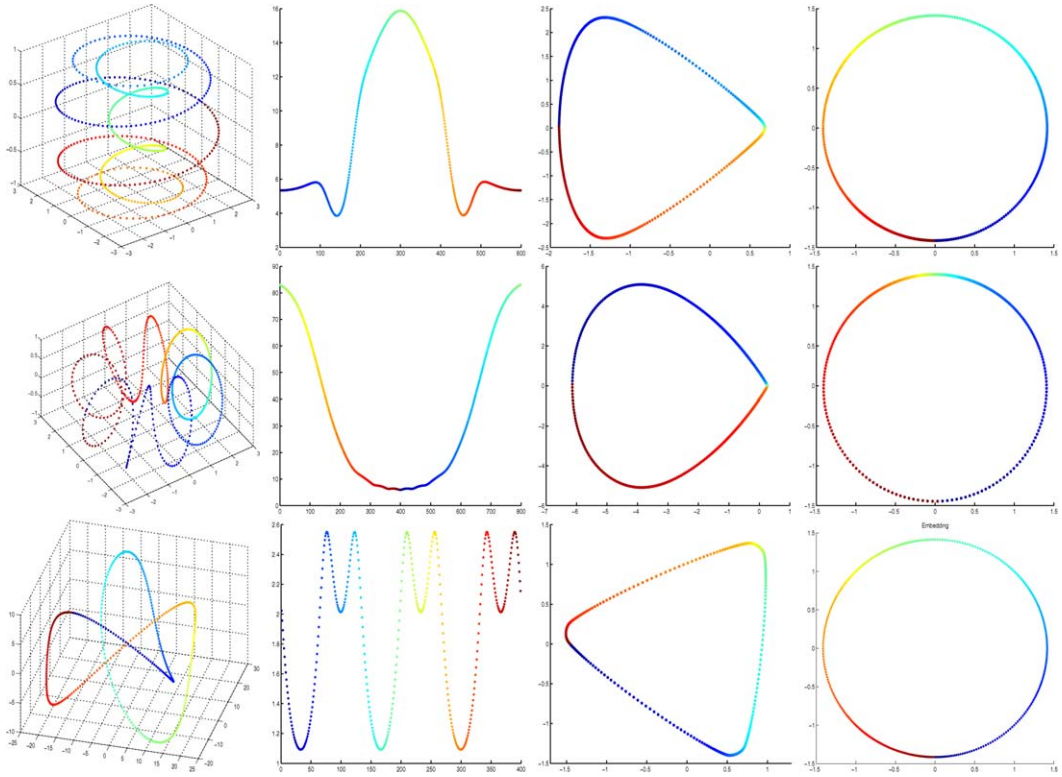


Fig. 4. From left to right: original curves, the densities of points, the embeddings via the graph Laplacian ($\alpha = 0$) and the embeddings via the Laplace–Beltrami approximation ($\alpha = 1$). In the latter case, the curve is embedded as a perfect circle and the arclength parametrization is recovered.

tribution of the data from the geometry of the underlying manifold and constitutes a simple way to approximate the Neumann heat kernel on $\mathcal{M}$.

Note that in the case of a general α, we also have that $P_{\varepsilon,1}^{\frac{t}{\varepsilon}}$ is an approximation of the diffusion kernel $\exp(-t(\Delta - \frac{\Delta q^{1-\alpha}}{q^{1-\alpha}}))$ of the corresponding stochastic process.

As an illustration, consider a set of points $X \in \mathbb{R}^3$ on a curve, with some nonuniform density. Although a natural ordering of these points is given by following the curve, the points were given *unordered*. We then computed and plotted the embedding obtained using the first 2 nontrivial eigenfunctions (ψ_1, ψ_2) of the Laplace–Beltrami operator, ignoring $\psi_0 = 1$. The results are shown in Fig. 4 for three different curves.

This experiment shows that the approximate Laplace–Beltrami diffusion (approximated by our algorithm for $\alpha = 1$) allows to reorganize the points on a circle, regardless of the density of points on the original curve. On this circle, the density as a function of arclength is the same as on the original curve: the geometry and the statistics of the points have been completely decoupled. On the contrary, the graph Laplacian (corresponding to the situation $\alpha = 0$) tends to generate corners at regions of high density. To understand the situation for the Laplace–Beltrami embedding, it is useful to recall the formula for the heat kernel at time t on a curve of length 1:

$$e_t(x, y) = 1 + \sum_{j \geqslant 1} e^{-j^2 t} \cos\bigl(2\pi j(\theta_x - \theta_y)\bigr),$$

where θ_x and θ_y are the arclength parameters of points x and y on the curve. This implies that the diffusion distance D_t at time t is verified:

$$e^t D_t^2(x, y) = \bigl|e^{2i\pi\theta_x} - e^{2i\pi\theta_y}\bigr|^2\bigl(1 + \mathcal{O}\bigl(e^{-3t}\bigr)\bigr)$$

and therefore, using $\psi_1(x) = \cos(2i\pi\theta_x)$ and $\psi_2(x) = \sin(2i\pi\theta_x)$, the curve is embedded as a circle of same length, and the diffusion distance is, at moderate and large times t, proportional to the cord-length between the images of x and y.

4. Directed diffusions

The previous section emphasizes the importance of anisotropic diffusions. We now extend this idea to the construction of kernels that define fast and slow directions for the Markov chain, in order to capture geometric features of interest or to perform a specific local analysis (see [16] for more details). We illustrate this idea by considering an empirical function f defined on X. Suppose that we want to find a function g defined on X that is constant on the level sets of f, but that is less oscillatory than f.

One way to proceed is to consider the kernel

$$k_\varepsilon(x, y) = \exp\left(-\frac{\|x - y\|^2}{\varepsilon} - \frac{(\langle \nabla f, x - y\rangle)^2}{\varepsilon^2}\right).$$

This kernel defines the affinity between points in X, and obviously, it will favor associations of points belonging to the same level set. It can be shown that by normalizing this kernel into a Markov kernel p_ε, then as $\varepsilon \to 0$, one obtains a diffusion along the levels sets of f (see [33]). In addition, the first nontrivial eigenfunction ψ_1 is constant along those level sets and has a few oscillations (since it is at the beginning of the spectrum of the Laplacian).

We illustrate these ideas by considering the function $f(\varphi, \theta) = \sin(12\theta)$ on the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ (see Fig. 5). This function oscillates 12 times along the z-axis, and by employing an f-driven diffusion, we were able to generate the function the function $\psi_1(\varphi, \theta) = \cos(\theta)$ as the first nontrivial eigenfunction.

5. Finite data set approximating $\mathcal{M}$

We now investigate the fact that in applications one has to deal with finite data sets and that even if the manifold assumption is valid, the approximation of $\mathcal{M}$ by the data points only holds modulo some perturbation due to noise or imperfections of the model on the data. A natural way to compute the different quantities involved in our algorithm is to approximate integrals by finite sums:

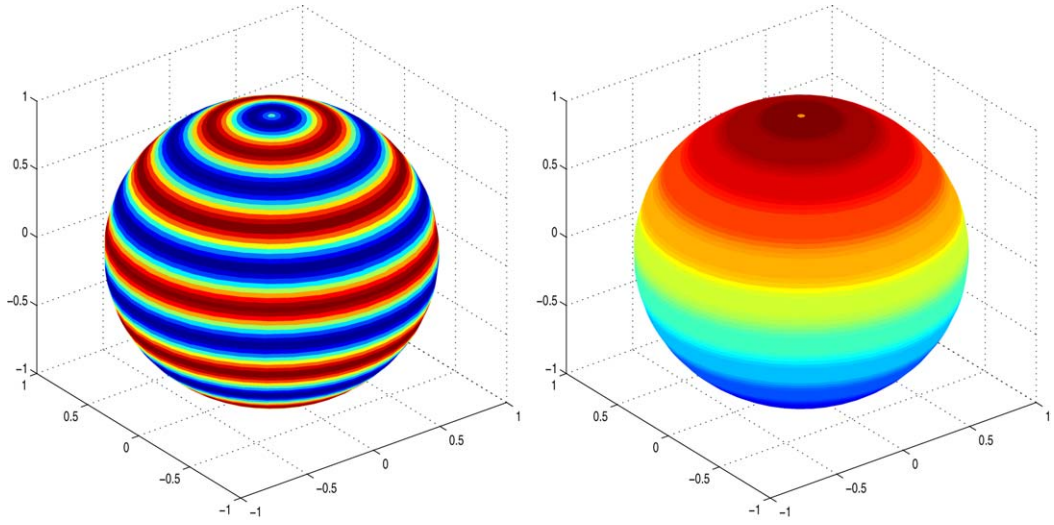


Fig. 5. Left: the original function $f(\varphi,\theta)=\sin(12\theta)$. Right: first nontrivial eigenfunction $\phi_1(\varphi,\theta)=\cos(\theta)$.

$$\bar{q}_\varepsilon(x_i)=\sum_{j=1}^{m}k_\varepsilon(x_i,x_j)\quad\text{and}\quad \bar{d}_\varepsilon^{(\alpha)}(x_i)=\sum_{j=1}^{m}\frac{k_\varepsilon(x_i,x_j)}{q_\varepsilon(x_i)^\alpha q_\varepsilon(x_j)^\alpha},$$

$$\bar{p}_{\varepsilon,\alpha}(x_i,x_j)=\frac{k_\varepsilon(x_i,x_j)}{d_\varepsilon^{(\alpha)}(x_i)d_\varepsilon^{(\alpha)}(x_j)}\quad\text{and}\quad \bar{P}_{\varepsilon,\alpha}f(x_i)=\sum_{j=1}^{m}p_\varepsilon(x_i,x_j)f(x_j),$$

where the bar over the letters means that these quantities represent discrete sums used for approximating their continuous counterparts, namely $q_\varepsilon(x_i)$, $d_\varepsilon^{(\alpha)}(x_i)$, $p_{\varepsilon,\alpha}(x_i,x_j)$ and $\bar{P}_{\varepsilon,\alpha}f(x_i)$ introduced in the previous sections.

In our study, two factors of error are to be accounted for: the fact that the data set is finite, and the fact that the points do not exactly lie on the manifold. We investigate the influence of these two factors separately.

Assume first that the data set $X=\{x_i\}_{1\leqslant i\leqslant m}$ consists of finitely many points on $\mathcal{M}$ that are the realizations of i.i.d. random variables $\{X_i\}_{1\leqslant i\leqslant m}$ with density q (supported on $\mathcal{M}$). Because of the law of large numbers, as m goes to infinity, all of the discrete sums above converge, at least in some weak sense and modulo a renormalization by $1/m$, to continuous integrals (Monte Carlo integration). For instance,

$$\lim_{m\to+\infty}\frac{1}{m}\sum_{j=1}^{m}k_\varepsilon(x,x_j)=\int_{\mathcal{M}}k_\varepsilon(x,y)q(y)\,\mathrm{d}y.$$

For a finite value of m, the relative error is expected to be of the order of $\mathcal{O}(m^{-\frac{1}{2}}\varepsilon^{-d/4})$ and the same estimate should apply to the error of approximating $P_{\varepsilon,\alpha}f(x_i)$ by $\bar{P}_{\varepsilon,\alpha}f(x_i)$.

Several papers provide rigorous estimates for the accuracy of the approximation. For instance, it is shown in [23] that the error of approximation of $P_{\varepsilon,\alpha}f(x_i)$ by $\bar{P}_{\varepsilon,\alpha}f(x_i)$ verifies

$$\left|\bar{P}_{\varepsilon,\alpha}f(x_i)-P_{\varepsilon,\alpha}f(x_i)\right|=\mathcal{O}\left(m^{-\frac{1}{2}}\varepsilon^{-d/4}\right),$$

with high probability.

This bound can be further refined [24] by noticing that the numerator and denominator of the expression defining $\bar{P}_{\varepsilon,\alpha}f(x_i)$ are correlated random variables. This allows Singer to derive the following estimate:

$$\left|\bar{P}_{\varepsilon,\alpha}f(x_i)-P_{\varepsilon,\alpha}f(x_i)\right|=\mathcal{O}\left(m^{-\frac{1}{2}}\varepsilon^{-d/4+1/2}\right),$$

with high probability.

Since we are interested in the approximation of $L_{\varepsilon,\alpha}$ by $\bar{L}_{\varepsilon,\alpha}$, we conclude the following: with high probability

$$\left|\bar{L}_{\varepsilon,\alpha}f(x_i)-L_{\varepsilon,\alpha}f(x_i)\right|=\mathcal{O}\left(m^{-\frac{1}{2}}\varepsilon^{-d/4-1/2}\right).$$

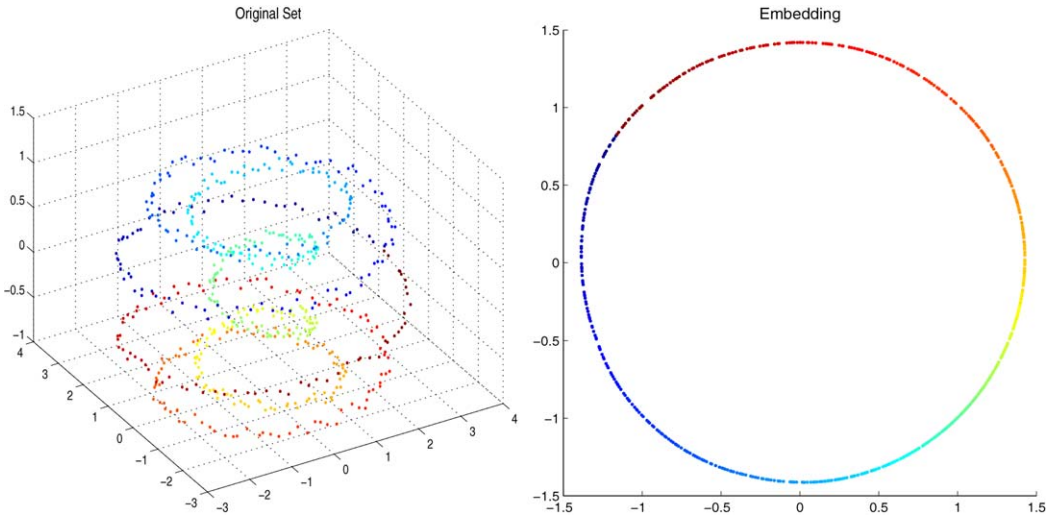


Fig. 6. Left: a noisy helix. Right: the curve is embedded as a perfect circle.

Criterion 4. In order to achieve a given precision with high probability, the number m of sample points must grow faster than $\varepsilon^{-\frac{d}{4}-\frac{1}{2}}$, where d is the dimension of $\mathcal{M}$.

The second aspect that we wish to explore concerns the fact that the data points of X might not lie exactly on $\mathcal{M}$. This case can be treated using spectral perturbation theory. Suppose that X is a perturbed version of $\mathcal{M}$, that is, there exists a perturbation function $\eta : \mathcal{M} \to X$ with a small norm (the size of the perturbation) such that every point in X can be written as $x + \eta(x)$ for some $x \in \mathcal{M}$. The function η plays the role of some additive noise on the data. Then, assuming that the kernel k_ε is smooth, we can linearize the effect of the perturbation:

$$k_\varepsilon\big(x + \eta(x), y + \eta(y)\big) = k_\varepsilon(x, y) + \mathcal{O}\left(\frac{\|\eta\|}{\sqrt{\varepsilon}}\right)$$

and as a consequence, the perturbation of $P_{\varepsilon,\alpha}$ is the same order

$$\tilde{P}_{\varepsilon,\alpha} = P_{\varepsilon,\alpha} + \mathcal{O}\left(\frac{\|\eta\|}{\sqrt{\varepsilon}}\right), \tag{3}$$

where $\tilde{P}_{\varepsilon,\alpha}$ is the perturbed version of $P_{\varepsilon,\alpha}$. To obtain the effect of the perturbation on the eigenvalues and eigenfunctions, we refer to classical theorems of spectral perturbation theory, like Weyl's theorem, that states that

$$\sup_l |\tilde{\lambda}_l - \lambda_l| \leqslant \|\tilde{P}_{\varepsilon,\alpha} - P_{\varepsilon,\alpha}\|.$$

The bound on the error in Eq. (3) shows that

Criterion 5. The approximation is valid as long as the scale parameter $\sqrt{\varepsilon}$ remains larger than the size of the perturbation.

We now illustrate the robustness of the different objects we have defined so far (diffusion distances, maps and eigenfunctions) to perturbations. In Fig. 6, we considered a helix curve in $\mathbb{R}^3$ that we perturbed with some additive Gaussian noise. We represented the embedding obtained from (ψ_1, ψ_2) and the image by the diffusion map is a perfect circle. Compared to the organization of the data points that we would have obtained from a perfect helix, this organization is identical up to the scale defined by the standard deviation of the noise.

6. Combining the multiscale geometries

The collection $\{D_t\}_{t \geqslant 0}$ describes the data set X through a family of multiscale geometries corresponding to the diffusion at different time-scales. The descriptions provided by the diffusion maps $\{\Psi_t\}_{t \geqslant 0}$ represents the data set at

different resolutions. In applications, one can deal with the different values of m separately, or one can combine the information at all times into one single description. For instance, in [19], Fouss et al. suggest the use of the *average commute time* $c(x, y)$ as a distance between points of X. The quantity $c(x, y)$ is defined as the average minimum time for the random walker to go from x to y and then return to x, in other words, $c(x, y) = m(x, y) + m(y, x)$, where $m(x, y)$ is the average first time passage to y when starting at x. The average commute time and first time passage exhibit the same interesting property as the diffusion distance D_t, namely that they get smaller as the number of paths joining x and y increases. As shown in [19], these quantities can be computed using the pseudo-inverse of the combinatorial graph Laplacian, and this leads to another type of embedding of the data set into a Euclidean space, in which the usual distance equals the average commute time, and where the balls of diffusion are specified by the level sets of the Green's function.

In order to illustrate the usefulness of combining diffusions at all times, we focus on the example of partially labeled data classification. Suppose that one is given a data set X, made of m points, and that one wants to classify all these points among r classes. Assume that within the set X, there are s points already classified, but this number is very small compared to the size of X. This type of situation is quite common in many applications, and goes by the name of partially-labeled data. In [2], Szummer and Jaakkola suggest a classification scheme based on random walks on the data, which basically comes down to computing the posterior probability for a point to belong to a given class, at a given time t. The value of t is chosen as maximizing the margin between the classes. An alternative way to classify the data consists in considering the labeled points as absorbing states and to use the average time of absorption (i.e., the average first time passage in a labeled point) as a classifier: x is assigned to the same class as the labeled point y that is the closest in terms of average first time passage. Another option would be to use the probability of absorption by a given absorbing state as a classifier:

$$\text{class}(x) = \arg\max_j P(\text{getting absorbed by class } j \mid \text{starting at } x).$$

The probability of absorption by a given state for a random walk starting at x is generally referred to as the Harmonic measure ω_x, and this approach is equivalent to considering the absorbing states as part of a "boundary" and solving the following Dirichlet problem using Kakutani's theorem: find the harmonic function (in the sense defined by the random walk) whose restriction to the boundary is equal to the characteristic function of all points with a given label.

7. Conclusion

In this paper, we have introduced a set of tools, the diffusion maps and distances, and we have explained their construction from Markov processes defined on a set X. The diffusion coordinates provide a global representation of X via a coordinate system that integrate the local geometry of the graph of the data. These coordinates allow us to parametrize the data set, and they also define a metric on X that reflects the connectivity within the graph. Dimensionality reduction is achieved thanks to the decay of the eigenvalues of the diffusion operator.

In the case of data points in the Euclidean space $\mathbb{R}^n$, we constructed a one-parameter family of diffusion kernels that capture different features of the data. In particular, we showed how to design a kernel that reproduces the diffusion induced by a Fokker–Planck equation. This allows to compute the long-time dynamics of some stochastic differential system. Also, we explained that when the data approximate a manifold, then one can recover the geometry of this manifold by computing an approximation of the Laplace–Beltrami operator. This computation is completely insensitive to the distribution of the points and therefore provides a separation of the statistics and the geometry of the data.

In the future, we plan to show that working in the diffusion space can be extremely useful for applications related to pattern recognition, learning, multisensor integration and the study of dynamical systems.

Acknowledgments

We thank Ann Lee, Mauro Maggioni, Boaz Nadler, Naoki Saito and Amit Singer for useful discussions regarding this work. We also thank Matthias Hein for his comments. Last, we are grateful to the reviewers for their helpful suggestions.

Appendix A. Spectral decomposition of P and diffusion distance

If we conjugate p by $\sqrt{\pi}$, then the resulting kernel

$$a(x,y) = \frac{\sqrt{\pi(x)}}{\sqrt{\pi(y)}} p(x,y) = \frac{k(x,y)}{\sqrt{\pi(x)}\sqrt{\pi(y)}}$$

is symmetric. The corresponding operator A is therefore self-adjoint in $L^2(X, \mathrm{d}\mu)$. If, in addition, we assume that

$$\int_X \int_X a(x,y)^2 \,\mathrm{d}\mu(y)\,\mathrm{d}\mu(x) = \int_X \int_X k(x,y)^2 \frac{\mathrm{d}\mu(y)}{\pi(y)} \frac{\mathrm{d}\mu(x)}{\pi(y)} < +\infty,$$

then A is also compact (see [32, p. 94] for a reference). Note that this condition is always satisfied if X is finite as A can be represented as a finite matrix. From the compactness of P, we have that A has a *discrete* set of eigenvalues $\{\lambda_l\}_{l\geqslant 0}$ and that it satisfies the following eigendecomposition:

$$a(x,y) = \sum_{l\geqslant 0} \lambda_l \phi_l(x)\phi_l(y),$$

where $\{\phi_l\}_{l\geqslant 0}$ is an orthonormal set of eigenfunctions forming a basis of $L^2(X, \mathrm{d}\mu)$. It can be checked that $\phi_0 = \sqrt{\pi}$.

This implies that p satisfies

$$p(x,y) = \sum_{l\geqslant 0} \lambda_l \psi_l(x)\varphi_l(y),$$

where $\psi_l(x) = \phi_l(x)/\pi(x)$ and $\varphi_l(y) = \phi_l(y)\pi(y)$. In particular, $\psi_0(x) = 1$. Moreover, by a result [33, p. 24] the largest eigenvalue λ_0 of P is equal to 1. As discussed previously, it is reasonable to assume that $k(x,x) > 0$ as this number represents this affinity of x with itself. As a consequence, the Markov chain is aperiodic. The graph being connected, it is also irreducible. These two conditions imply that the chain is ergodic and that, apart from λ_0, all other eigenvalues have a magnitude stricly less than 1.

We have an analogous formula for powers P^t of P:

$$p_t(x,y) = \sum_{l\geqslant 0} \lambda_l^t \psi_l(x)\varphi_l(y). \tag{A.1}$$

Since $\{\phi_l\}_{l\geqslant 0}$ forms an orthonormal basis of $L^2(X, \mathrm{d}\mu)$, it can be verified that $\{\varphi_l\}_{l\geqslant 0}$ is an orthonormal basis of $L^2(X, \mathrm{d}\mu/\pi)$.

The connection with diffusion distances is as follows: if we fix x, then Eq. (A.1) can be viewed as the orthogonal expansion of the function $y \mapsto p_t(x,y)$ into the orthonormal basis $\{\varphi_l\}_{l\geqslant 0}$. Note that the coefficients of expansion are precisely given by the sequence of numbers $\{\lambda_l^t \psi_l(x)\}_{l\geqslant 0}$. Consequently, in $L^2(X, \mathrm{d}\mu/\pi)$:

$$D_t(x,z)^2 = \left\| p_t(x,\cdot) - p_t(z,\cdot) \right\|^2_{L^2(X,\mathrm{d}\mu/\pi)} = \sum_{l\geqslant 0} \lambda_l^{2t} \big(\psi_l(x) - \psi_l(z)\big)^2.$$

Appendix B. Asymptotics for Laplacian operators

In this appendix, we prove different asymptotic expansions for the various operators defined in Section 3 and acting on a submanifold $\mathcal{M} \subset \mathbb{R}^n$. We extend the proofs of [33] from the case of hypersurfaces to general submanifolds. We start by comparing the metric on the manifold (given by the geodesic distance) with that defined on the (local) projection of the submanifold over the tangent space. We use this result to obtain a small-scale expansion for any isotropic kernel integral operator. Then we conclude with the asymptotics for all the normalizations introduced in Section 3.

We assume $\mathcal{M}$ to be C^∞ and compact. Let $x \in \mathcal{M}$ be a fixed point not on the boundary, $T_x\mathcal{M}$ be the tangent space to $\mathcal{M}$ at x and $(e_1, \ldots, e_d)$ be a fixed orthonormal basis of $T_x\mathcal{M}$. In what follows, we introduce two systems of local coordinates in the neighborhood of x (see Fig. B.1). We first consider normal coordinates: the exponential map $\exp_x$ generates a set of orthogonal geodesics $(\gamma_1, \ldots, \gamma_d)$ intersecting at x with initial velocity $(e_1, \ldots, e_d)$, and any

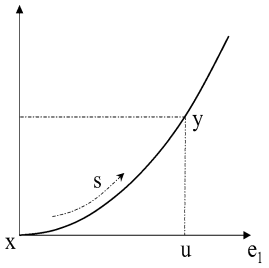


Fig. B.1. The two types of coordinates that we use, illustrated here for a curve in the 2D plane.

point y in a sufficiently small neighborhood of x has a set of normal coordinates $(s_1, \ldots, s_d)$ along these geodesics. Therefore, any function f defined on $\mathcal{M}$ in the neighborhood of x, can be viewed as a function $\tilde{f}$ of $(s_1, \ldots, s_d)$, and in this case, if f is $\mathcal{C}^2$, we have

$$\Delta f(x) = -\sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0, \ldots, 0),$$

where Δ is the Laplace–Beltrami operator on $\mathcal{M}$ (see [30]). With this convention, Δ is a positive semi-definite operator. If x is on the boundary $\partial\mathcal{M}$ of $\mathcal{M}$, and if we choose $e_1, \ldots, e_{d-1}$ to be in the tangent space of this boundary at x, while e_d is normal and pointing in, then the normal derivative of a function f at x is defined as

$$\frac{\partial f}{\partial \nu}(x) = -\frac{\partial \tilde{f}}{\partial s_d}(0).$$

The other system of coordinates is given by the orthogonal projection u of y on $T_x\mathcal{M}$. More precisely, let $(u_1, \ldots, u_d)$ be its coordinates in $(e_1, \ldots, e_d)$, i.e., $u_i = \langle y - x, e_i \rangle$. The submanifold is now locally parameterized as $y = (u, g(u))$, where $g : \mathbb{R}^d \to \mathbb{R}^{n-d}$. Note that since $u = (u_1, \ldots, u_d)$ are tangent coordinates, we must have that $\frac{\partial g}{\partial u_i}(0) = 0$.

Locally, we have the following diagram:

$$(s_1, \ldots, s_d) \overset{\exp_x}{\longleftrightarrow} y \xleftarrow{\text{projection}} u.$$

In what follows, we provide the tools for converting all quantities depending on $(s_1, \ldots, s_d)$ or y into functions of u. More precisely, we compute asymptotic expansions for the changes of variable $u \mapsto (s_1, \ldots, s_d)$ and $u \mapsto y$. Notice that similar results were obtained in [31], where the geodesic distance is compared to the Euclidean distance in the ambient space $\mathbb{R}^n$ (rather than the projection on the tangent space) and in [12] for the asymptotics of the graph Laplacian for uniform densities on submanifolds. In the following, $Q_{x,m}(u)$ denotes a generic homogeneous polynomial of degree m of the variable $u = (u_1, \ldots, u_d)$, whose coefficient depends on x. Since these polynomials form an equivalence class, we might abuse the notation and write, for instance, $Q_{x,m}(u) + Q_{x,m}(u) = Q_{x,m}(u)$.

Lemma 6. *If $y \in \mathcal{M}$ is in a Euclidean ball of radius $\varepsilon^{\frac{1}{2}}$ around x, then, for ε sufficiently small, there exists*:

$$s_i = u_i + Q_{x,3}(u) + \mathcal{O}\big(\varepsilon^2\big). \tag{B.1}$$

Proof. Let γ be the geodesic connecting x and y parameterized by arclength. We have $\gamma(0) = x$ and let s be such that $\gamma(s) = y$. If y has normal coordinates $(s_1, \ldots, s_d)$, then we have $s\gamma'(0) = (s_1, \ldots, s_d)$. A Taylor expansion yields

$$\gamma(s) = \gamma(0) + s\gamma'(0) + \frac{s^2}{2}\gamma''(0) + \frac{s^3}{6}\gamma^{(3)}(0) + \mathcal{O}\big(\varepsilon^2\big).$$

By definition of a geodesic, the covariant derivative of the velocity is zero, which means that $\gamma''(0)$ is orthogonal to the tangent plane at x. Now since the parameter u_i is defined by $u_i = \langle \gamma(s) - \gamma(0), e_i \rangle$, we obtain that $u_i = s_i + \frac{s^3}{6}\langle \gamma^{(3)}(0), e_i \rangle + \mathcal{O}(\varepsilon^2)$. Iterating this equation yields the result. □

Lemma 7. *In the same ball as in the previous lemma, we have*

$$\|y-x\|^2 = \|u\|^2 + Q_{x,4}(u) + Q_{x,5}(u) + \mathcal{O}\big(\varepsilon^3\big) \quad \text{(metric comparison)} \tag{B.2}$$

and

$$\det\left(\frac{\mathrm{d}y}{\mathrm{d}u}\right) = 1 + Q_{x,2}(u) + Q_{x,3}(u) + \mathcal{O}\big(\varepsilon^2\big) \quad \text{(volume comparison)}. \tag{B.3}$$

Proof. The submanifold is locally parameterized as $u \mapsto (u, g(u))$, where $g : \mathbb{R}^d \to \mathbb{R}^{n-d}$. Writing $g = (g_{i+1}, \ldots, g_n)$ and applying Pythagore's theorem, we obtain

$$\|y-x\|^2 = \|u\|^2 + \sum_{i=d+1}^{n} g_i(u)^2.$$

Clearly, $g_i(0) = 0$, and as noted before, $\frac{\partial g}{\partial u_i}(0) = 0$. As a consequence, $g_i(u) = b_{i,x}(u) + c_{i,x}(u) + \mathcal{O}(\varepsilon^2)$, where $b_{i,x}$ is the Hessian quadratic form of g_i at $u = 0$ and $c_{i,x}$ is the cubic term. This proves (B.2) with

$$Q_{x,4}(u) = \sum_{i=d+1}^{n} b_{i,x}^2(u) \quad \text{and} \quad Q_{x,5}(u) = 2\sum_{i=d+1}^{n} b_{i,x}(u)c_{i,x}(u).$$

To prove Eq. (B.3), observe that the fact that $\frac{\partial g}{\partial u_i}(0) = 0$ implies that $\frac{\partial g}{\partial u_i}(u) = \tilde{b}_{i,x}(u) + \tilde{c}_{i,x}(u) + \mathcal{O}(\varepsilon^{\frac{3}{2}})$, where $\tilde{b}_{i,x}$ and $\tilde{c}_{i,x}(u)+$ are the linear and quadratic terms in the Taylor expansion of $\frac{\partial g}{\partial u_i}$ at 0. We thus have:

$$\begin{aligned}\frac{\partial y}{\partial u_i}(u) &= \left(v_i, \frac{\partial g}{\partial u_i}(u)\right), \quad \text{where } v_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^d \\ &= \big(v_i, \tilde{b}_{i,x}(u) + \tilde{c}_{i,x}(u) + \mathcal{O}\big(\varepsilon^{\frac{3}{2}}\big)\big).\end{aligned}$$

The squared volume generated by these d vectors is the determinant of their Gram matrix, i.e.,

$$\left|\det\left(\frac{\mathrm{d}y}{\mathrm{d}u}\right)\right|^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} E_{ij}(u) + \sum_{i=1}^{d}\sum_{j=1}^{d} F_{ij}(u) + \mathcal{O}\big(\varepsilon^2\big),$$

where

$$E_{ij}(u) = \big\langle \tilde{b}_{i,x}(u), \tilde{b}_{j,x}(u)\big\rangle \quad \text{and} \quad F_{ij}(u) = \big\langle \tilde{b}_{i,x}(u), \tilde{c}_{j,x}(u)\big\rangle + \big\langle \tilde{c}_{i,x}(u), \tilde{b}_{j,x}(u)\big\rangle.$$

Defining

$$Q_{x,2}(u) = \sum_{i=1}^{d}\sum_{j=1}^{d} E_{ij}(u) \quad \text{and} \quad Q_{x,3}(u) = \sum_{i=1}^{d}\sum_{j=1}^{d} F_{ij}(u),$$

we obtain the last result. □

Let $k_\varepsilon(x, y)$ be an isotropic kernel:

$$k_\varepsilon(x, y) = h\left(\frac{\|x-y\|^2}{\varepsilon}\right),$$

where h is assumed to have an exponential decay and let G_ε be the corresponding operator acting on

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}}\int_{\mathcal{M}} k_\varepsilon(x, y) f(y)\,\mathrm{d}y.$$

We now produce asymptotics for G_ε. The idea is to use the previous lemmas to write that, for small values of the scale parameter ε, integrating a function f against the kernel on the manifold is approximately like integrating on the tangent space. On this space, the kernel is approximately symmetric and is orthogonal to monomials of degree 1 and therefore

will only capture terms of degree 0 and 2 in the Taylor expansion of f. All corrections to these approximations are expressed in terms of extrinsic and intrinsic geometric quantities related to $\mathcal{M}$. We start by studying the case of points away from the boundary.

Lemma 8. *Let $f \in \mathcal{C}^3(\mathcal{M})$ and let $0 < \gamma < 1/2$. Then we have, uniformly for all $x \in \mathcal{M}$ at distance larger than ε^γ from $\partial\mathcal{M}$,*

$$G_\varepsilon f(x) = m_0 f(x) + \varepsilon \frac{m_2}{2}\big(\omega(x) f(x) - \Delta f(x)\big) + \mathcal{O}\big(\varepsilon^2\big),$$

where

$$m_0 = \int_{\mathbb{R}^d} h\big(\|u\|^2\big)\,\mathrm{d}u \quad \text{and} \quad m_2 = \int_{\mathbb{R}^d} u_1^2 h\big(\|u\|^2\big)\,\mathrm{d}u$$

and ω is a potential term depending on the embedding of $\mathcal{M}$.

Proof. Because of the exponential decay of h, the domain of integration can be restricted to the intersection of $\mathcal{M}$ with the ball of radius ε^γ around x. Indeed, in doing so, we generate an error of order

$$\begin{aligned}\left|\frac{1}{\varepsilon^{\frac{d}{2}}} \int_{y\in\mathcal{M}:\|y-x\|>\varepsilon^\gamma} h\left(\frac{\|x-y\|^2}{\varepsilon}\right) f(y)\,\mathrm{d}y\right| &\leqslant \|f\|_\infty \frac{1}{\varepsilon^{\frac{d}{2}}} \int_{y\in\mathcal{M}:\|y-x\|>\varepsilon^\gamma} \left|h\left(\frac{\|x-y\|^2}{\varepsilon}\right)\right| \mathrm{d}y \\ &\leqslant \|f\|_\infty \int_{y\in\mathcal{M}:\|y\|>\varepsilon^{\gamma-1/2}} \big|h\big(\|y\|^2\big)\big|\,\mathrm{d}y \\ &\leqslant C\|f\|_\infty Q\big(\varepsilon^{1/2-\gamma}\big) e^{-\varepsilon^{\gamma-1/2}},\end{aligned}$$

where we have used the exponential decay of the kernel and where Q is a polynomial. Since $0 < \gamma < 1/2$, this term is exponentially small and is bounded by $\mathcal{O}(\varepsilon^{\frac{3}{2}})$. Therefore,

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}} \int_{y\in\mathcal{M}:\|y-x\|<\varepsilon^\gamma} h\left(\frac{\|x-y\|^2}{\varepsilon}\right) f(y)\,\mathrm{d}y + \mathcal{O}\big(\varepsilon^{\frac{3}{2}}\big).$$

Now that things are localized around x, we can Taylor-expand the function $(s_1,\ldots,s_d) \mapsto f(y(s_1,\ldots,s_d))$:

$$f(y) = f(x) + \sum_{i=1}^d s_i \frac{\partial \tilde{f}}{\partial s_i}(0) + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d s_i s_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j}(0) + Q_{x,3}(s_1,\ldots,s_d) + \mathcal{O}\big(\varepsilon^2\big),$$

where $\tilde{f}(s_1,\ldots,s_d) = f(y(s_1,\ldots,s_d))$. Invoking Eq. (B.1), we obtain

$$f(y) = \tilde{f}(0) + \sum_{i=1}^d u_i \frac{\partial \tilde{f}}{\partial s_i}(0) + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d u_i u_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j}(0) + Q_{x,3}(u) + \mathcal{O}\big(\varepsilon^2\big).$$

Likewise, because of Eq. (B.2), the Taylor expansion of the kernel is

$$h\left(\frac{\|x-y\|^2}{\varepsilon}\right) = h\left(\frac{\|u\|^2}{\varepsilon}\right) + \left(\frac{Q_{x,4}(u)}{\varepsilon} + \frac{Q_{x,5}(u)}{\varepsilon}\right) h'\left(\frac{\|u\|^2}{\varepsilon}\right) + \mathcal{O}\big(\varepsilon^2\big).$$

Using Eq. (B.3) to change the variable $s \mapsto u$ in the previous integral defining $G_\varepsilon f(x)$ yields:

$$\begin{aligned}\varepsilon^{\frac{d}{2}} G_\varepsilon f(x) = \int_{\|u\|<\varepsilon^\gamma} &\left(h\left(\frac{\|u\|^2}{\varepsilon}\right) + \left(\frac{Q_{x,4}(u)}{\varepsilon} + \frac{Q_{x,5}(u)}{\varepsilon}\right) h'\left(\frac{\|u\|^2}{\varepsilon}\right)\right) \\ &\times \left(\tilde{f}(0) + \sum_{i=1}^d u_i \frac{\partial \tilde{f}}{\partial s_i}(0) + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d u_i u_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j}(0) + Q_{x,3}(u)\right) \\ &\times \big(1 + Q_{x,2}(u) + Q_{x,3}(u)\big)\,\mathrm{d}u + \mathcal{O}\big(\varepsilon^{\frac{d}{2}+2}\big).\end{aligned}$$

This identity can be dramatically simplified by identifying odd functions and setting their integral to zero. One is left with

$$\varepsilon^{\frac{d}{2}} G_\varepsilon f(x) = \tilde{f}(0) \int_{\mathbb{R}^d} h\left(\frac{\|u\|^2}{\varepsilon}\right) \mathrm{d}u + \frac{1}{2}\left(\sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0)\right) \int_{\mathbb{R}^d} u_1^2 h\left(\frac{\|u\|^2}{\varepsilon}\right) \mathrm{d}u + \tilde{f}(0) \int_{\mathbb{R}^d} \left(\frac{Q_{x,4}(u)}{\varepsilon} h'\left(\frac{\|u\|^2}{\varepsilon}\right) + \tilde{Q}_{x,2}(u) h\left(\frac{\|u\|^2}{\varepsilon}\right)\right) \mathrm{d}u + \mathcal{O}\left(\varepsilon^{\frac{d}{2}+2}\right),$$

where the domain of integration has been extended to $\mathbb{R}^d$ (exponential decay of h). Changing the variable according to $u \mapsto \sqrt{\varepsilon} u$,

$$G_\varepsilon f(x) = \tilde{f}(0) \int_{\mathbb{R}^d} h\left(\|u\|^2\right) \mathrm{d}u + \frac{\varepsilon}{2}\left(\sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0)\right) \int_{\mathbb{R}^d} u_1^2 h\left(\|u\|^2\right) \mathrm{d}u + \varepsilon \tilde{f}(0) \int_{\mathbb{R}^d} \left(Q_{x,4}(u) h'\left(\|u\|^2\right) + Q_{x,2}(u) h\left(\|u\|^2\right)\right) \mathrm{d}u + \mathcal{O}\left(\varepsilon^2\right),$$

where we have used the homogeneity of $Q_{x,4}$ and $Q_{x,2}$. Finally, observing that

$$\tilde{f}(0) = f(x) \quad \text{and} \quad \sum_{i=1}^{d} \frac{\partial^2 \tilde{f}}{\partial s_i^2}(0) = -\Delta f(x),$$

we end up with

$$G_\varepsilon f(x) = m_0 f(x) + \varepsilon \frac{m_2}{2}\left(\omega(x) f(x) - \Delta f(x)\right) + \mathcal{O}\left(\varepsilon^2\right),$$

where

$$\omega(x) = \frac{2}{m_2} \int_{\mathbb{R}^d} \left(Q_{x,4}(u) h'\left(\|u\|^2\right) + Q_{x,2}(u) h\left(\|u\|^2\right)\right) \mathrm{d}u.$$

Finally, the uniformity follows from the compactness and smoothness of $\mathcal{M}$. □

The case of points close to the boundary is a bit more delicate. Indeed, if we integrate a function f against the kernel centered at such a point, the kernel will "see" the first-order term of the Taylor series of f (because it does not correspond to a vanishing moment anymore). In fact the only first term that is captured by the kernel is the derivative across the boundary since in other directions, the kernel is symmetric, and therefore orthogonal to monomials of degree 1.

Lemma 9. *Let $f \in \mathcal{C}^3(\mathcal{M})$ and let $0 < \gamma \leqslant 1/2$. Then we have, uniformly for all $x \in \mathcal{M}$ at distance less than or equal to ε^γ from $\partial\mathcal{M}$,*

$$G_\varepsilon f(x) = m_0^\varepsilon(x) f(x_0) + \sqrt{\varepsilon} m_1^\varepsilon(x) \frac{\partial f}{\partial \nu}(x_0) + \mathcal{O}(\varepsilon),$$

where x_0 is the closest point to x that belongs to the boundary and where $m_0^\varepsilon(x)$ and $m_1^\varepsilon(x)$ are bounded functions of x and ε.

Proof. Let x_0 be the closest point to x on the boundary, where closeness is measured with the norm of the ambient space. This point is uniquely defined if the boundary is smooth and ε is small enough. Let us pick a specific orthonormal basis of $T_{x_0}\mathcal{M}$ so that the first $(d-1)$ vectors $e_1, \ldots, e_{d-1}$ belong to the tangent space $T_{x_0}\partial\mathcal{M}$ of the boundary at x_0 (see Fig. B.2). As before, we can now consider the projections $u = (v, u_d)$ of points y in the neighborhood of x onto $T_{x_0}\mathcal{M}$, where $v = (u_1, \ldots, u_{d-1}) \in \mathbb{R}^{d-1}$ is the projection over the first $(d-1)$ basis vectors and $u_d \in \mathbb{R}$ is the projection over e_d (pointing in). By definition of x_0, we have $\langle x - x_0, e_i \rangle = 0$ for $i = 1, \ldots, d-1$, and therefore x has

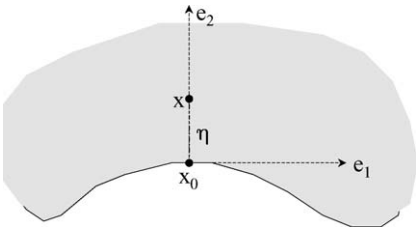


Fig. B.2. Geometric configuration at the boundary.

coordinates $(0, \eta)$ where $\eta \geqslant 0$. The proof is very similar to that of the previous lemma. Indeed, up to a term of order ε, the integrand is symmetric with respect to the variable u_i for $i = 1, \ldots, d-1$, and thus the integration will "kill" the corresponding first order terms in the Taylor expansion of $\tilde{f}$. The only term that remains is the normal derivative.

Just as before, we can truncate the integral defining $G_\varepsilon f(x)$ by considering only points y that are at most at distance ε^γ from x. The correction term is exponentially small, and therefore can be bounded by $\mathcal{O}(\varepsilon)$. In addition to this truncation, we decompose the domain into slices. More precisely, define

$$S(u_d) = \left\{(v, u_d) \in \mathbb{R}^d \middle/ \left\|(v, u_d) - (0, \eta)\right\| \leqslant \varepsilon^\gamma\right\}.$$

To compute the integral defining $G_\varepsilon f(x)$ up to order ε, we can integrate over all $S(u_d)$ for $u_d \in [\eta - \varepsilon^\gamma; \eta + \varepsilon^\gamma]$. Now this is not good enough as we want to take advantage of the symmetries of the kernel. We therefore consider

$$\tilde{S}(u_d) = \bigcap_{i=1}^{d-1} R_i S(u_d),$$

where R_i is the reflection on $\mathbb{R}^d$ defined by

$$R_i(u_1, \ldots, u_{i-1}, u_i, u_{i+1}, \ldots, u_d) = (u_1, \ldots, u_{i-1}, -u_i, u_{i+1}, \ldots, u_d).$$

This domain has now all the symmetries that we need. Moreover, up to a term of order $\varepsilon^{\frac{3}{2}}$, the projection of $\partial\mathcal{M}$ onto $T_{x_0}\mathcal{M}$ is a hypersurface (in $\mathbb{R}^d$) with equation $u_d = \varphi(u_1, \ldots, u_{d-1})$, where φ is a homogeneous polynomial of degree 2. Consequently, up to an error of the same order, it is approximately preserved by all reflections R_i. In particular, going from the slices $S(u_d)$ to $\tilde{S}(u_d)$ is only generating an error of order ε.

$$G_\varepsilon f(x) = \varepsilon^{-\frac{d}{2}} \int_{\eta-\varepsilon^\gamma}^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) \tilde{f}(u)\, \mathrm{d}v\, \mathrm{d}u_d + \mathcal{O}(\varepsilon),$$

where $u = (v, u_d)$. For the same reason, starting the integration from $u_d = 0$ generates an error of order ε:

$$G_\varepsilon f(x) = \varepsilon^{-\frac{d}{2}} \int_{0}^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) \tilde{f}(u)\, \mathrm{d}v\, \mathrm{d}u_d + \mathcal{O}(\varepsilon).$$

If we Taylor-expand $\tilde{f}$ around $u = 0$, we obtain

$$\tilde{f}(u) = \tilde{f}(0) + \sum_{i=1}^{d} u_i \frac{\partial \tilde{f}}{\partial u_i}(0) + \mathcal{O}(\varepsilon) = f(x_0) + \sum_{i=1}^{d-1} u_i \frac{\partial \tilde{f}}{\partial u_i}(0) - u_d \frac{\partial f}{\partial v}(x_0) + \mathcal{O}(\varepsilon).$$

Now the symmetry of the kernel implies that for $i = 1, \ldots, d-1$,

$$\int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) u_i\, \mathrm{d}v = 0.$$

Therefore, the only first-order term of the Taylor expansion that survives is the partial derivative along u_d. We can conclude that

$$G_\varepsilon f(x) = m_0^\varepsilon(x) f(x_0) + \sqrt{\varepsilon} m_1^\varepsilon(x) \frac{\partial f}{\partial v}(x_0) + \mathcal{O}(\varepsilon),$$

with

$$m_0^\varepsilon(x_0) = \varepsilon^{-\frac{d}{2}} \int_0^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) \mathrm{d}v \, \mathrm{d}u_d = \int_{-\eta/\sqrt{\varepsilon}}^{\varepsilon^{\gamma-1/2}} \int_{\frac{1}{\sqrt{\varepsilon}}\tilde{S}(u_d)} h\big(\|u\|^2\big) \mathrm{d}v \, \mathrm{d}u_d$$

and

$$m_1^\varepsilon(x_0) = -\varepsilon^{-\frac{d}{2}} \int_0^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) \frac{u_d}{\sqrt{\varepsilon}} \mathrm{d}v \, \mathrm{d}u_d = -\int_{-\eta/\sqrt{\varepsilon}}^{\varepsilon^{\gamma-1/2}} u_d \int_{\frac{1}{\sqrt{\varepsilon}}\tilde{S}(u_d)} h\big(\|u\|^2\big) \mathrm{d}v \, \mathrm{d}u_d.$$

Clearly, these functions are well behaved as

$$\big|m_0^\varepsilon(x)\big| \leqslant \int_{\mathbb{R}^d} h\big(\|u\|^2\big) \mathrm{d}u \quad \text{and} \quad \big|m_1^\varepsilon(x)\big| \leqslant \int_{\mathbb{R}^d} \big|u_d h\big(\|u\|^2\big)\big| \mathrm{d}u.$$

The uniformity follows from the compactness and smoothness of $\mathcal{M}$ and of its boundary ∂M. □

We now use these ingredients to obtain asymptotics for operator $L_{\varepsilon,\alpha}$. To simplify the notations, we can assume that function h in Lemma 8 is scaled in such a way that $m_0 = 1$ and $m_2 = 2$. We now prove

Proposition 10. *For a fixed $K > 0$, we have on E_K*

$$\lim_{\varepsilon \to 0} L_{\varepsilon,\alpha} f = \frac{\Delta(f q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}} f.$$

Proof. We fix $0 < \gamma < 1/2$ and we start by focusing on the set $\mathcal{M}_\varepsilon$ of points of $\mathcal{M}$ that are at distance larger than ε^γ from $\partial\mathcal{M}$. Using the notations introduced in Section 3 and Lemma 8, it is easy to verify that uniformly on $\mathcal{M}_\varepsilon$ and up to a term of order $\mathcal{O}(\varepsilon^2)$:

$$q_\varepsilon = q + \varepsilon(\omega q - \Delta q) \tag{B.4}$$

and that, consequently,

$$q_\varepsilon^{-\alpha} = q^{-\alpha}\left(1 + \alpha\varepsilon\left(\frac{\Delta q}{q} - \omega\right)\right). \tag{B.5}$$

Let

$$k_\varepsilon^{(\alpha)}(x, y) = \frac{k_\varepsilon(x, y)}{q_\varepsilon^\alpha(x) q_\varepsilon^\alpha(y)} \tag{B.6}$$

and for $\phi \in E_K$, define

$$K_\varepsilon^{(\alpha)} \phi(x) = \int_X k_\varepsilon^{(\alpha)}(x, y) \phi(y) q(y) \, \mathrm{d}y. \tag{B.7}$$

Then

$$\begin{aligned} K_\varepsilon^{(\alpha)} \phi &= q_\varepsilon^{-\alpha}\left[\phi q^{1-\alpha}\left(1 + \varepsilon\alpha\left(\frac{\Delta q}{q} - \omega\right)\right) + \varepsilon\big(\omega\phi q^{1-\alpha} - \Delta\big(\phi q^{1-\alpha}\big)\big)\right] \\ &= q_\varepsilon^{-\alpha}\left[\phi q^{1-\alpha} + \varepsilon\left((1-\alpha)\omega\phi q^{1-\alpha} + \alpha\phi\frac{\Delta q}{q^\alpha} - \Delta\big(\phi q^{1-\alpha}\big)\right)\right] \\ &= q_\varepsilon^{-\alpha} q^{1-\alpha}\left[\phi + \varepsilon\left((1-\alpha)\omega\phi + \alpha\phi\frac{\Delta q}{q} - \frac{\Delta(\phi q^{1-\alpha})}{q^{1-\alpha}}\right)\right]. \end{aligned}$$

Consequently,

$$d_\varepsilon^{(\alpha)} = K_\varepsilon^{(\alpha)}.1 = q_\varepsilon^{-\alpha} q^{1-\alpha}\left[1 + \varepsilon\left((1-\alpha)\omega + \alpha\frac{\Delta q}{q} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\right)\right]. \tag{B.8}$$

Taking the ratio of the last two equations yields the expansion for the operator

$$\int_X \frac{k_\varepsilon^{(\alpha)}(x,y)}{d_\varepsilon^{(\alpha)}(x)}\phi(y)q(y)\,\mathrm{d}y = \phi(x) + \varepsilon\left(\phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)} - \frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)}\right). \tag{B.9}$$

Therefore, uniformly on $\mathcal{M}_\varepsilon$,

$$L_{\varepsilon,\alpha}\phi = \frac{\Delta(\phi q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\phi + \mathcal{O}(\varepsilon). \tag{B.10}$$

Now, on $\mathcal{M}\backslash\mathcal{M}_\varepsilon$, Lemma 9 implies that, uniformly and up to a term of order ε,

$$q_\varepsilon(x) = m_0^\varepsilon(x)q(x_0) + \sqrt{\varepsilon}m_1^\varepsilon(x)\frac{\partial q}{\partial \nu}(x_0) + \mathcal{O}(\varepsilon),$$

which implies that

$$q_\varepsilon^{-\alpha}(x) = m_0^\varepsilon(x)^{-\alpha} q(x_0)^{-\alpha}\left(1 - \alpha\sqrt{\varepsilon}\frac{m_1^\varepsilon(x)}{m_0^\varepsilon(x)}\frac{1}{q(x_0)}\frac{\partial q}{\partial \nu}(x_0)\right) + \mathcal{O}(\varepsilon).$$

As a consequence,

$$\begin{aligned}
K_\varepsilon^{(\alpha)}\phi(x) &= q_\varepsilon^{-\alpha}(x_0)\left(\frac{q(x_0)^{1-\alpha}}{m_0^\varepsilon(x)^{\alpha-1}}\phi(x_0) + \sqrt{\varepsilon}m_1^\varepsilon(x)\frac{\partial\left(q^{1-\alpha}(m_0^\varepsilon)^{-\alpha}\phi\right)}{\partial\nu}(x_0)\right.\\
&\qquad \left. - \alpha\sqrt{\varepsilon}m_0^\varepsilon(x)^{-\alpha}q(x_0)^{1-\alpha}m_1^\varepsilon(x)\frac{1}{q(x_0)}\frac{\partial q}{\partial\nu}(x_0)\phi(x_0)\right) + \mathcal{O}(\varepsilon)\\
&= q_\varepsilon^{-\alpha}(x_0)\left(\frac{q(x_0)^{1-\alpha}}{m_0^\varepsilon(x)^{\alpha-1}}\phi(x_0) + \sqrt{\varepsilon}m_1^\varepsilon(x)\frac{\partial\left(q^{1-\alpha}(m_0^\varepsilon)^{-\alpha}\right)}{\partial\nu}(x_0)\phi(x_0)\right.\\
&\qquad \left. - \alpha\sqrt{\varepsilon}m_0^\varepsilon(x)^{-\alpha}m_1^\varepsilon(x)\frac{1}{q(x_0)^\alpha}\frac{\partial q}{\partial\nu}(x_0)\phi(x_0)\right) + \mathcal{O}(\varepsilon),
\end{aligned}$$

where we have used the fact that ϕ verifies the Neumann condition at x_0 and therefore can be taken out of any derivative across the boundary. Thus,

$$K_\varepsilon^{(\alpha)}\phi(x) = \left(K_\varepsilon^{(\alpha)}.1 + \mathcal{O}(\varepsilon)\right)\phi(x_0)$$

and since $d_\varepsilon^{(\alpha)} = K_\varepsilon^{(\alpha)}.1$, for $x \in \mathcal{M}\backslash\mathcal{M}_\varepsilon$,

$$\frac{K_\varepsilon^{(\alpha)}\phi(x)}{d_\varepsilon^{(\alpha)}} = \phi(x_0) + \mathcal{O}(\varepsilon)$$

and, therefore, uniformly on $\mathcal{M}\backslash\mathcal{M}_\varepsilon$,

$$L_{\varepsilon,\alpha}\phi(x) = \mathcal{O}(1). \tag{B.11}$$

To summarize the situation:

- Uniformly on $\mathcal{M}_\varepsilon$, we have

$$L_{\varepsilon,\alpha}\phi = \frac{\Delta(\phi q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\phi + \mathcal{O}(\varepsilon),$$

- and uniformly on $\mathcal{M}\backslash\mathcal{M}_\varepsilon$, we have $L_{\varepsilon,\alpha}\phi(x) = \mathcal{O}(1)$.

Since we are interested in the L^2 convergence of this operator on $\mathcal{M}$, and for the strip $\mathcal{M}\backslash\mathcal{M}_\varepsilon$ has a measure of order $\mathcal{O}(\varepsilon^\gamma)$, we have on E_K:

$$L_{\varepsilon,\alpha}\phi = \frac{\Delta(\phi q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\phi + R_\varepsilon,$$

where $R_\varepsilon = \mathcal{O}(\varepsilon)$ if $\mathcal{M}$ has no boundary and $R_\varepsilon = \mathcal{O}(\varepsilon^\gamma)$ for any $\gamma \in (0; \frac{1}{2})$ if $\mathcal{M}$ has a boundary. □

We finally prove Proposition 3:

Proposition 11. *For any $t > 0$, the Neumann heat kernel $e^{-t\Delta}$ can be approximated on $L^2(\mathcal{M})$ by $P_{\varepsilon,1}^{\frac{t}{\varepsilon}}$:*

$$\lim_{\varepsilon\to 0} P_{\varepsilon,1}^{\frac{t}{\varepsilon}} = e^{-t\Delta}.$$

Proof. The idea of the proof is that at very small scales, $\mathcal{M}$ is flat and the Markov chain is a good approximation of the short-time transitions of heat. We then show that the semi-group property holds approximately, which allows us to propagate the approximation to large times. It is worth noting that this result is valid for general values of α, where one just needs to replace Δ by $\Delta - \frac{\Delta(p^{1-\alpha})}{p^{1-\alpha}}$. In this case, the following proof can be transposed easily.

In the previous proof, we showed that on E_K,

$$L_{\varepsilon,1} = \Delta + R_\varepsilon \quad \text{or equivalently} \quad P_\varepsilon = I - \varepsilon\Delta - \varepsilon R_\varepsilon.$$

To obtain the result on the heat kernel, we note that

- $\overline{\bigcup_{K>0} E_K} = L^2(\mathcal{M})$,
- $(P_{\varepsilon,1})_{\varepsilon>0}$ is uniformly bounded in $L^2(\mathcal{M}, dx)$ by 1

and therefore the result needs only to be proven on E_K for any fixed value of $K > 0$. We also remark that if B is a bounded operator with $\|B\| < 1$ and nonnegative spectrum, then for any $\beta > 0$

$$\begin{aligned}\big\|(I+B)^\beta - I\big\| &= \left\| \sum_{l\geqslant 1} \frac{\beta(\beta-1)\cdots(\beta-l+1)}{l!} B^l \right\| \leqslant \sum_{l\geqslant 1} \frac{\beta(\beta-1)\cdots(\beta-l+1)}{l!} \|B\|^l \\ &\leqslant \big(1+\|B\|\big)^\beta - 1 \leqslant \beta\big(1+\|B\|\big)^{\beta-1}\|B\|. \end{aligned} \tag{B.12}$$

For a fixed $K > 0$, if ε is small enough, then $I - \varepsilon\Delta$ is invertible, and has norm less than 1, in which case

$$\begin{aligned}\big\|P_{\varepsilon,1}^{\frac{t}{\varepsilon}} - (I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}\big\| &\leqslant \big\|(I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}\big\| \big\|(I-\varepsilon\Delta)^{-\frac{t}{\varepsilon}}(I-\varepsilon\Delta-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - I\big\| \\ &\leqslant \big\|(I-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - I\big\| \\ &\leqslant \frac{t}{\varepsilon} \times \big(1+\varepsilon\|R_\varepsilon\|\big)^{\frac{t}{\varepsilon}-1} \times \varepsilon\|R_\varepsilon\| \quad \text{by Eq. (B.12)} \\ &= \mathcal{O}\big(\|R_\varepsilon\|\big).\end{aligned}$$

Now since on E_K, one has $(I-\varepsilon\Delta)^{\frac{t}{\varepsilon}} = e^{-t\Delta} + \mathcal{O}(\varepsilon)$ (this can be verified by viewing each operator in the eigenbasis of Δ), we can conclude that

$$P_{\varepsilon,1}^{\frac{t}{\varepsilon}} = e^{-t\Delta} + \mathcal{O}\big(\|R_\varepsilon\|\big). \qquad \square$$

References

[1] P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent space approaches to social networks analysis, J. Amer. Statist. Assoc. 97 (2002) 1090–1098.
[2] M. Szummer, T. Jaakkola, Partially labeled classification with Markov random walks, Neural Inf. Process. System 14 (2001) 945–952.
[3] F. Chung, Spectral Graph Theory, in: CBNS-AMS, vol. 92, Amer. Math. Soc., Providence, RI.
[4] J. Shi, J. Malik, Normalized cuts and image segmentation, in: Proc. IEEE Conf. CVPR, 1997, pp. 731–737.

[5] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1–7) (1998) 107–117.
[6] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical report, Stanford University, 1998.
[7] S. White, P. Smyth, Algorithms for estimating relative importance in networks, in: Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, ACM Press, 2003, pp. 266–275.
[8] T. Haveliwala, Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search, IEEE Trans. Knowl. Data Eng. 15 (4) (2003) 784–796.
[9] J. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.
[10] R. Lempel, S. Moran, The stochastic approach for link-structure analysis (SALSA) and the TKC effect, in: Proc. 9th Int. World Wide Web Conf., 2000, pp. 387–401.
[11] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 13 (2003) 1373–1397.
[12] M. Belkin, Problems of learning on manifolds, Ph.D. dissertation, August 2003.
[13] M. Kac, Can one hear the shape of a drum? Part II, Amer. Math. Monthly 73 (4) (1966) 1–23.
[14] A. Nahmod, Geometry of operators and spectral analysis, Ph.D. dissertation, Yale University, October 1991.
[15] Y. Safarov, D. Vassiliev, The Asymptotic Distribution of Eigenvalues of Partial Differential Operators, Amer. Math. Soc., Providence, RI, 1996.
[16] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data, Proc. Natl. Acad. Sci. (2004), submitted for publication.
[17] R.R. Coifman, M. Maggioni, Diffusion wavelets, Appl. Comput. Harmon. Anal. (2004), in press.
[18] D.L. Donoho, C. Grimes, Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. 100 (2003) 5591–5596.
[19] F. Fouss, A. Pirotte, J.-M. Renders, The application of new concepts of dissimilarities between nodes of a graph to collaborative filtering, 2004, submitted for publication.
[20] J. Ham, D.D. Lee, S. Mika, B. Schölkopf, A kernel view of the dimensionality reduction of manifolds, Technical report TR-110, Max Planck Institute for Biological Cybernectics, July 2003.
[21] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[22] E.M. Stein, Topics in harmonic analysis related to the Littlewood–Paley theory, Ann. Math. Stud., Princeton Univ. Press, Princeton, NJ, 1970.
[23] M. Hein, J. Audibert, U. vonLuxburg, From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians, in: Conference On Learning Theory, June 2005, pp. 470–485.
[24] A. Singer, From graph to manifold Laplacian: The convergence rate, Technical report, Yale University, 2005.
[25] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignement, Technical report CSE-02-019, Pennsylvania State University, 2002.
[26] Y. Weiss, Segmentation using eigenvectors: A unifying view, in: Proc. IEEE Int. Conf. on Computer Vision, 1999, pp. 975–982.
[27] K.S. Pedersen, A.B. Lee, Toward a full probability model of edges in natural images, in: M. Nielsen, A. Heyden, G. Sparr, P. Johansen (Eds.), Proc. 7th European Conference on Computer Vision, Springer, Copenhagen, 2002, pp. 328–342. Part I.
[28] P.S. Huggins, S.W. Zucker, Representing edge models via local principal component analysis, in: M. Nielsen, A. Heyden, G. Sparr, P. Johansen (Eds.), Proc. 7th European Conference on Computer Vision, Springer, Copenhagen, 2002, pp. 384–398. Part I.
[29] B. Nadler, S. Lafon, R.R. Coifman, Diffusion maps, spectral clustering and reaction coordinates of stochastic dynamical systems, Appl. Comput. Harmon. Anal. (2006), this issue.
[30] S. Rosenberg, The Laplacian on a Riemannian Manifold, Cambridge Univ. Press, 1997.
[31] O.G. Smolyanov, H.V. Weizsäcker, O. Wittich, Brownian motion on a manifold as limit of stepwise conditioned standard Brownian motions, Canad. Math. Soc. Conf. Proc. 29 (2000) 589–602.
[32] M. Pedersen, Functional Analysis in Applied Mathematics and Engineering, CRC Press, 1999.
[33] S. Lafon, Diffusion maps and geometric harmonics, Ph.D. dissertation, Yale University, 2004.