Available online at www.sciencedirect.com
SCIENCE DIRECT®
Appl. Comput. Harmon. Anal. 21 (2006) 113–127
Applied and Computational Harmonic Analysis
www.elsevier.com/locate/acha

Diffusion maps, spectral clustering and reaction coordinates of dynamical systems

Boaz Nadler [a,*], Stéphane Lafon [a,1], Ronald R. Coifman [a], Ioannis G. Kevrekidis [b]

[a] *Department of Mathematics, Yale University, New Haven, CT 06520, USA*
[b] *Chemical Engineering and PACM, Princeton University, Princeton, NJ 08544, USA*

Received 28 October 2004; revised 10 February 2005; accepted 29 July 2005

Available online 9 June 2006

Communicated by the Editors

Abstract

A central problem in data analysis is the low dimensional representation of high dimensional data and the concise description of its underlying geometry and density. In the analysis of large scale simulations of complex dynamical systems, where the notion of time evolution comes into play, important problems are the identification of slow variables and dynamically meaningful reaction coordinates that capture the long time evolution of the system. In this paper we provide a unifying view of these apparently different tasks, by considering a family of *diffusion maps*, defined as the embedding of complex (high dimensional) data onto a low dimensional Euclidean space, via the eigenvectors of suitably defined random walks defined on the given datasets. Assuming that the data is randomly sampled from an underlying general probability distribution $p(\mathbf{x}) = e^{-U(\mathbf{x})}$, we show that as the number of samples goes to infinity, the eigenvectors of each diffusion map converge to the eigenfunctions of a corresponding differential operator defined on the support of the probability distribution. Different normalizations of the Markov chain on the graph lead to different limiting differential operators. Specifically, the normalized graph Laplacian leads to a backward Fokker–Planck operator with an underlying potential of $2U(\mathbf{x})$, best suited for spectral clustering. A different anisotropic normalization of the random walk leads to the backward Fokker–Planck operator with the potential $U(\mathbf{x})$, best suited for the analysis of the long time asymptotics of high dimensional stochastic systems governed by a stochastic differential equation with the same potential $U(\mathbf{x})$. Finally, yet another normalization leads to the eigenfunctions of the Laplace–Beltrami (heat) operator on the manifold in which the data resides, best suited for the analysis of the geometry of the dataset regardless of its possibly non-uniform density.

1. Introduction

Analysis of complex high dimensional data is an exploding area of research, with applications in diverse fields, such as machine learning, statistical data analysis, bio-informatics, meteorology, chemistry and physics, to mention only a few. In the first three application fields, the underlying assumption is that the data is sampled from some unknown probability distribution, typically without any notion of time or correlation between consecutive samples. Important

* Corresponding author. Currently at: Weizmann Institute of Science, Rehovot, 76100, Israel.
E-mail addresses: boaz.nadler@weizmann.ac.il (B. Nadler), coifman@math.yale.edu (R.R. Coifman).
[1] Currently at Google.

1063-5203/$ – see front matter © 2006 Published by Elsevier Inc.
doi:10.1016/j.acha.2005.07.004

tasks are dimensionality reduction, e.g., representation of the high dimensional data with only a few coordinates and the study of the geometry and statistics of the data, its possible decomposition into clusters, etc. [1]. In addition, there are many problems concerning supervised and semi-supervised learning, in which additional information, such as a discrete class $g(\mathbf{x}) \in \{g_1, \ldots, g_k\}$ or a continuous function value $f(\mathbf{x})$ is known for some or all of the data points. In this paper we consider only the unsupervised case, although some of the methods and ideas presented can be applied to the supervised or semi-supervised cases as well [2].

In the later three above-mentioned application fields the data is typically sampled from a complex biological, chemical or physical *dynamical* system, in which there is an inherent notion of time. These systems typically involve multiple time and length scales, but in many interesting cases there is a separation of time scales, that is, there are only a few "slow" time scales at which the system performs structural changes from one meta-stable state to another, with many additional fast time scales at which the system performs local fluctuations within these meta-stable states. In the case of macromolecules the slow time scale is that of a conformational change, while the fast time scales are governed by the chaotic rotations and vibrations of the individual chemical bonds between the different atoms of the molecule, as well as the random fluctuations due to the frequent collisions with the surrounding solvent water molecules. In the more general case of interacting particle systems, the fast time scales are those of density fluctuations around the mean density profiles while the slow time scales correspond to the time evolution of these mean density profiles.

Although on the fine time and length scales the full description of such systems requires a high dimensional space, e.g., the locations (and velocities) of all the different particles, these systems typically have an intrinsic low dimensionality on coarser length and time scales. Thus, the coarse time evolution of the high dimensional system can be described by only a few dynamically relevant variables, typically called reaction coordinates. Important tasks in such systems are the reduction of the dimensionality at these coarser scales (known as homogenization) and the efficient representation of the complicated linear or non-linear operators that govern their (coarse grained) time evolution. Additional goals are identification of meta-stable states, characterization of the transitions between them and efficient computation of mean exit times, potentials of mean force and effective diffusion coefficients [3–6].

In this paper, following [7], we consider a family of diffusion maps for the analysis of these problems. Given a large dataset, we construct a family of random walk processes on the data based on isotropic and anisotropic diffusion kernels and study their first few eigenvalues and eigenvectors (principal components). The key point in our analysis is that these eigenvectors and eigenvalues capture important geometrical and statistical information regarding the structure of the underlying datasets.

It is interesting to note that similar approaches have been suggested in many different fields. Use of the second eigenvector of the graph Laplacian has a long history, dating back at least to Fidler's work in the 1970's [8]. In recent years the first few eigenvectors of the *normalized* graph Laplacian were suggested for spectral clustering, image segmentation and dimensionality reduction [9–13], while similar constructions have been used for clustering and identification of meta-stable states from simulations of dynamical systems [4]. On the theoretical front, in [12,14] Belkin and Niyogi showed that for data sampled uniformly from an underlying manifold, the first few eigenvectors are discrete approximations of the eigenfunctions of the Laplace–Beltrami operator on the manifold, thus providing a justification for their use as a dimensional reduction tool. A different analysis, based on the observation that the normalized graph Laplacian defines a random walk on the data was performed by various authors [15–17].

In this paper, we provide a unified probabilistic framework for these methods and consider in detail the connection of these eigenvalues and eigenvectors to the underlying geometry and probability density distribution of the dataset. To this end, we assume that the data is sampled from some (unknown) probability distribution and view the eigenvectors computed on the finite dataset as discrete approximations of corresponding eigenfunctions of suitably defined continuum operators in an infinite population setting. As the number of samples goes to infinity, the discrete random walk on the set converges to a diffusion process defined on the n-dimensional space but with a non-uniform probability density. By explicitly studying the asymptotic form of the Chapman–Kolmogorov equations in this setting (e.g., the infinitesimal generators), we find that for data sampled from a general probability distribution, written in Boltzmann form as $p(\mathbf{x}) = e^{-U(\mathbf{x})}$, the eigenvectors and eigenvalues of the standard normalized graph Laplacian construction correspond to a diffusion process with a potential $2U(\mathbf{x})$ (instead of a single $U(\mathbf{x})$). Therefore, a subset of the first few eigenfunctions are indeed well suited for clustering of data that contains only a few well separated clusters, corresponding to deep wells in the potential $U(\mathbf{x})$.

Motivated by the well-known connection between diffusion processes and Schrödinger operators [18], we propose a different novel non-isotropic construction of a random walk on the graph, that in the asymptotic limit of infinite data

recovers the eigenvalues and eigenfunctions of a diffusion process with the same potential $U(\mathbf{x})$. This normalization, therefore, is most suited for the study of the long time behavior of complex dynamical systems that evolve in time according to a stochastic differential equation. For example, in the case of a dynamical system driven by a bistable potential with two wells (e.g., with one slow time scale for the transition between the wells and many fast time scales), the second eigenfunction can serve as a parametrization of the reaction coordinate between the two states, much in analogy to its use as an approximation to the optimal normalized cut for graph segmentation. For the analysis of dynamical systems, we also propose to use a subset of the first few eigenfunctions as reaction coordinates for the design of fast simulations. The main idea is that once a parametrization of dynamically meaningful reaction coordinates is known and lifting and projection operators between the original space and the diffusion map are available, detailed simulations can be initialized at different locations on the reaction path and run only for short times, to estimate the transition probabilities to different nearby locations in the reaction coordinate space, thus efficiently constructing a potential of mean field and an effective diffusion coefficient on the reaction path [19].

Finally, we describe yet another random walk construction that in the limit of infinite data recovers the Laplace–Beltrami (heat) operator on the manifold on which the data resides, regardless of the possibly non-uniform sampling of points on it. This normalization is therefore best suited for learning the *geometry* of the dataset, as it separates the geometry of the manifold from the statistics on it.

Our analysis thus reveals the intimate connection between the eigenvalues and eigenvectors of different random walks on the finite graph to the underlying geometry and probability distribution from which the dataset was sampled. These findings lead to a better understanding of the characteristics, advantages and limitations of diffusion maps as a tool to solve different tasks in the analysis of high dimensional data.

2. Problem setup

Consider a finite dataset $\{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^n$ with two possible different scenarios for its origin. In the first scenario, the data is not necessarily derived from a dynamical system but rather it is randomly sampled from some arbitrary probability distribution $p(\mathbf{x})$ in a compact domain $\Omega \subset \mathbb{R}^n$. In this case we define an associated potential

$$U(\mathbf{x}) = -\log p(\mathbf{x}) \tag{1}$$

so that $p = e^{-U}$.

In the second scenario, we assume that the data is sampled from a dynamical system in equilibrium. We further assume that the dynamical system, defined by its state $\mathbf{x}(t) \in \Omega$ at time t, satisfies the following non-dimensional stochastic differential equation (SDE):

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

with reflecting boundary conditions on $\partial\Omega$, where a dot on a variable means differentiation with respect to time, U is the free energy at $\mathbf{x}$ (which, with some abuse of nomenclature, we will also call the potential at $\mathbf{x}$) and $\mathbf{w}(t)$ is n-dimensional Brownian motion. In this case there is an explicit notion of time and the transition probability density $p(\mathbf{x}, t \mid \mathbf{y}, s)$ of finding the system at location $\mathbf{x}$ at time t, given an initial location $\mathbf{y}$ at time s ($t > s$), satisfies the forward Fokker–Planck equation (FPE) [20,21]

$$\frac{\partial p}{\partial t} = \nabla \cdot \big(\nabla p + p \nabla U(\mathbf{x})\big) \tag{3}$$

with initial condition

$$\lim_{t \to s^+} p(\mathbf{x}, t \mid \mathbf{y}, s) = \delta(\mathbf{x} - \mathbf{y}). \tag{4}$$

Similarly, the backward Fokker–Planck equation for the density $p(\mathbf{x}, t \mid \mathbf{y}, s)$ in the backward variables $\mathbf{y}, s$ ($s < t$) is

$$-\frac{\partial p}{\partial s} = \Delta p - \nabla p \cdot \nabla U(\mathbf{y}) \tag{5}$$

where differentiations in (5) are with respect to the variable $\mathbf{y}$ and the Laplacian Δ is a negative operator, defined as $\Delta u = \nabla \cdot (\nabla u)$.

As time $t \to \infty$ the steady state solution of (3) is given by the equilibrium Boltzmann probability density,

$$\mu(\mathbf{x})\,d\mathbf{x} = \Pr\{\mathbf{x}\}\,d\mathbf{x} = \frac{\exp(-U(\mathbf{x}))}{Z}\,d\mathbf{x}, \tag{6}$$

where Z is a normalization constant (known as the partition function in statistical physics), given by

$$Z = \int_{\Omega} \exp\bigl(-U(\mathbf{x})\bigr)\,d\mathbf{x}. \tag{7}$$

In what follows we assume that the potential $U(\mathbf{x})$ is shifted by the suitable constant (which does not change the SDE (2)), so that $Z = 1$. Also, we use the notation $\mu(\mathbf{x}) = \Pr\{\mathbf{x}\} = p(\mathbf{x}) = e^{-U(\mathbf{x})}$ interchangeably to denote the (invariant) probability measure on the space.

Note that in both scenarios, the steady state probability density, given by (6) is identical. Therefore, for the purpose of our initial analysis, which does not directly take into account the possible time dependence of the data, it is only the features of the underlying potential $U(\mathbf{x})$ and the geometry of Ω that come into play.

The Langevin equation (2) or the corresponding Fokker–Planck equation (3) are commonly used to describe the time evolution of mechanical, physical, chemical, or biological systems driven by noise. The study of their behavior and specifically the decay to equilibrium has been the subject of much theoretical research [21,22]. In general, the solution of the Fokker–Planck equation (3) can be written in terms of an eigenfunction expansion

$$p(\mathbf{x}, t) = \sum_{j=0}^{\infty} a_j e^{-\lambda_j t} \varphi_j(\mathbf{x}), \tag{8}$$

where $-\lambda_j$ are the eigenvalues of the FP operator, with $\lambda_0 = 0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots$, $\varphi_j(\mathbf{x})$ are their corresponding eigenfunctions and the coefficients a_j depend on the initial conditions. Obviously, the long term behavior of the system is approximately governed by only the first few eigenfunctions $\varphi_0, \varphi_1, \ldots, \varphi_k$, where k is typically small and depends on the time scale of interest. In low dimensions, e.g., $n \leqslant 3$ for example, it is possible to numerically approximate these eigenfunctions via space discretization methods of the FP operator. In high dimensions, however, this approach is in general infeasible and one typically resorts to simulations of trajectories of the corresponding SDE (2). In this case, there is a need to employ statistical methods to analyze the simulated trajectories, identify slow variables, meta-stable states, reaction pathways connecting them and mean first passage times between them. As described in this paper, approximations to φ_j and to the coefficients a_j can be computed from a large set of simulated data.

3. Diffusion maps

3.1. Finite data

Let $\{\mathbf{x}_i\}_{i=1}^N$ denote N data samples, either merged from many different simulations of the stochastic equation (2) or simply given without an underlying dynamical system. In [7], Coifman and Lafon suggested the following method, based on the definition of a Markov chain on the data, for the analysis of the geometry of general datasets.

For a fixed value of ε (a metaparameter of the algorithm), define an isotropic diffusion kernel,

$$k_\varepsilon(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{4\varepsilon}\right). \tag{9}$$

Assume that the transition probability between points $\mathbf{x}_i$ and $\mathbf{x}_j$ is proportional to $k_\varepsilon(\mathbf{x}_i, \mathbf{x}_j)$ and construct an $N \times N$ Markov matrix as follows

$$M(i, j) = \frac{k_\varepsilon(\mathbf{x}_i, \mathbf{x}_j)}{p_\varepsilon(\mathbf{x}_i)}, \tag{10}$$

where p_ε is the required normalization constant, given by

$$p_\varepsilon(\mathbf{x}_i) = \sum_j k_\varepsilon(\mathbf{x}_i, \mathbf{x}_j). \tag{11}$$

For large enough values of ε the Markov matrix M is fully connected (in the numerical sense) and therefore has an eigenvalue $\lambda_0 = 1$ with multiplicity one and a sequence of additional $n-1$ non-increasing eigenvalues $\lambda_j < 1$, with corresponding right eigenvectors ψ_j.

The stochastic matrix M naturally induces a distance between any two data points, based on their dynamic proximity. Specifically, we define a *diffusion distance* at time t as follows

$$D_t^2(\mathbf{x},\mathbf{y}) = \left\| p(\mathbf{z},t \mid \mathbf{x}) - p(\mathbf{z},t \mid \mathbf{y}) \right\|_w^2 = \sum_{\mathbf{z}} \big(p(\mathbf{z},t \mid \mathbf{x}) - p(\mathbf{z},t \mid \mathbf{y})\big)^2 w(\mathbf{z}),$$

where $p(\mathbf{z},t \mid \mathbf{x})$ is the probability that the random walk is located at $\mathbf{z}$ at time t given a starting location $\mathbf{x}$ at time $t=0$. As shown in [23], with the weight function $w(\mathbf{z}) = 1/p_\varepsilon(\mathbf{z})$, we have the following identity:

$$D_t^2(\mathbf{x},\mathbf{y}) = \sum_j \lambda_j^{2t} \big(\psi_j(\mathbf{x}) - \psi_j(\mathbf{y})\big)^2. \tag{12}$$

We thus define the *diffusion map* at time m as the mapping from $\mathbf{x}$ to the vector

$$\Psi_m(\mathbf{x}) = \big(\lambda_0^m \psi_0(\mathbf{x}), \lambda_1^m \psi_1(\mathbf{x}), \ldots, \lambda_k^m \psi_k(\mathbf{x})\big)$$

for some small value k. According to (12), for large enough k the Euclidean distance between the diffusion map coordinates is approximately equal to the diffusion distance between the points in the original space.

In [7], it was demonstrated that this mapping gives a low dimensional parametrization of the geometry and density of the data. In the field of data analysis, this construction is known as the *normalized graph Laplacian*. In [11], Shi and Malik suggested using the first non-trivial eigenvector to compute an approximation to the optimal normalized cut of a graph, while the first few eigenvectors were suggested by Weiss et al. [9,10] for clustering. Similar constructions, falling under the general term of kernel methods have been used in the machine learning community for classification and regression [24]. In this paper we elucidate the connection between this construction, the structure of the eigenvectors and eigenvalues and the underlying potential $U(\mathbf{x})$ and geometry of Ω.

3.2. The limiting diffusion process

To analyze the eigenvalues and eigenvectors of the normalized graph Laplacian, we consider them as a finite approximation of a suitably defined diffusion operator acting on the continuous probability space from which the data was sampled. We thus consider the limit of the above Markov chain process as the number of samples approaches infinity. For a finite value of ε, the Markov chain in discrete time and space converges to a Markov process in discrete time but continuous space. Then, in the limit $\varepsilon \to 0$, this jump process converges to a diffusion process in Ω, whose local transition probability depends on the non-uniform probability measure $\mu(\mathbf{x}) = e^{-U(\mathbf{x})}$.

We first consider the case of a fixed $\varepsilon > 0$ and take $N \to \infty$. Using the similarity of (9) to the diffusion kernel, we view ε as a measure of time and consider a discrete jump process at time intervals $\Delta t = \varepsilon$, with a transition probability between points $\mathbf{y}$ and $\mathbf{x}$ proportional to $k_\varepsilon(\mathbf{x},\mathbf{y})$. However, since the density of points is not uniform but rather given by the measure $\mu(\mathbf{x})$, we define an associated normalization factor $p_\varepsilon(\mathbf{y})$ as follows

$$p_\varepsilon(\mathbf{y}) = \int_\Omega k_\varepsilon(\mathbf{x},\mathbf{y})\mu(\mathbf{x})\,d\mathbf{x} \tag{13}$$

and a forward transition probability

$$M_f(\mathbf{x} \mid \mathbf{y}) = \Pr\big(\mathbf{x}(t+\varepsilon) = \mathbf{x} \mid \mathbf{x}(t) = \mathbf{y}\big) = \frac{k_\varepsilon(\mathbf{x},\mathbf{y})}{p_\varepsilon(\mathbf{y})}. \tag{14}$$

Equations (13) and (14) are the continuous analogues of the discrete equations (11) and (10). For future use, we also define a symmetric kernel $M_s(\mathbf{x},\mathbf{y})$ as follows:

$$M_s(\mathbf{x},\mathbf{y}) = \frac{k_\varepsilon(\mathbf{x},\mathbf{y})}{\sqrt{p_\varepsilon(\mathbf{x})p_\varepsilon(\mathbf{y})}}. \tag{15}$$

Note that $p_\varepsilon(\mathbf{x})$ is an estimate of the local probability density at $\mathbf{x}$, computed by averaging the density in a neighborhood of radius $O(\varepsilon^{1/2})$ around $\mathbf{x}$. Indeed, for a unit normalized kernel, as $\varepsilon \to 0$ we have that

$$p_\varepsilon(\mathbf{x}) = p(\mathbf{x}) + \varepsilon \Delta p(\mathbf{x}) + O\big(\varepsilon^{3/2}\big). \tag{16}$$

We now define forward, backward and symmetric Chapman–Kolmogorov operators on functions defined on this probability space as follows:

$$T_f[\varphi](\mathbf{x}) = \int_{\Omega} M_f(\mathbf{x} \mid \mathbf{y})\varphi(\mathbf{y})\,\mathrm{d}\mu(\mathbf{y}), \tag{17}$$

$$T_b[\psi](\mathbf{x}) = \int_{\Omega} M_f(\mathbf{y} \mid \mathbf{x})\psi(\mathbf{y})\,\mathrm{d}\mu(\mathbf{y}) \tag{18}$$

and

$$T_s[\varphi](\mathbf{x}) = \int_{\Omega} M_s(\mathbf{x}, \mathbf{y})\varphi(\mathbf{y})\,\mathrm{d}\mu(\mathbf{y}). \tag{19}$$

If $\varphi(\mathbf{x})$ is the probability of finding the system at location $\mathbf{x}$ at time $t = 0$, then $T_f[\varphi]$ is the evolution of this probability to time $t = \varepsilon$. Similarly, if $\psi(\mathbf{z})$ is some function on the space, then $T_b[\psi](\mathbf{x})$ is the mean (average) value of that function at time ε for a random walk that started at $\mathbf{x}$ and so $T_b^m[\psi](\mathbf{x})$ is the average value of the function at time $t = m\varepsilon$.

By definition, the operators T_f and T_b are adjoint under the inner product with weight μ, while the operator T_s is self adjoint under this inner product,

$$\langle T_f\varphi, \psi\rangle_\mu = \langle \varphi, T_b\psi\rangle_\mu, \qquad \langle T_s\varphi, \psi\rangle_\mu = \langle \varphi, T_s\psi\rangle_\mu. \tag{20}$$

Moreover, since T_s is obtained via conjugation of the kernel M_f with $\sqrt{p_\varepsilon(\mathbf{x})}$ all three operators share the same eigenvalues. The corresponding eigenfunctions can be found via conjugation by $\sqrt{p_\varepsilon}$. For example, if $T_s\varphi_s = \lambda\varphi_s$, then the corresponding eigenfunctions for T_f and T_b are $\varphi_f = \sqrt{p_\varepsilon}\varphi_s$ and $\varphi_b = \varphi_s/\sqrt{p_\varepsilon}$, respectively. Since $\sqrt{p_\varepsilon}$ is the first eigenfunction with $\lambda_0 = 1$ of T_s, the steady state of the forward operator is simply $p_\varepsilon(\mathbf{x})$, while for the backward operator it is the constant function $\psi_b = 1$.

Obviously, the eigenvalues and eigenvectors of the discrete Markov chain described in the previous section are discrete approximations to the eigenvalues and eigenfunctions of these continuum operators. Mathematical proofs of this convergence as $N \to \infty$ under various assumptions appear in [7,14,28,29]. Therefore, for a better understanding of the finite sample case, we are interested in the properties of the eigenvalues and eigenfunctions of either one of the operators T_f, T_b or T_s, and how these relate to the measure $\mu(\mathbf{x})$ (or equivalently to corresponding potential $U(\mathbf{x})$) and to the geometry of Ω. To this end, we look for functions $\varphi(\mathbf{x})$ such that

$$T_j\varphi = \int_{\Omega} M_j(\mathbf{x}, \mathbf{y})\varphi(\mathbf{y})\Pr\{\mathbf{y}\}\,\mathrm{d}\mathbf{y} = \lambda\varphi(\mathbf{x}), \tag{21}$$

where $j \in \{f, b, s\}$.

While in the case of a finite amount of data, ε must remain finite so as to have enough neighbors in a ball of radius $O(\varepsilon^{1/2})$ near each point $\mathbf{x}$, as the number of samples tends to infinity we can take smaller and smaller values of ε. Therefore, it is instructive to look at the limit $\varepsilon \to 0$. In this case, the transition probability densities of the continuous in space but discrete in time Markov chain converge to those of a diffusion process, whose time evolution is described by a differential equation

$$\frac{\partial p}{\partial t} = \mathcal{H}_f p,$$

where $\mathcal{H}_f$ is the infinitesimal generator or propagator of the forward operator, defined as

$$\mathcal{H}_f = \lim_{\varepsilon\to 0} \frac{T_f - I}{\varepsilon}.$$

As shown in Appendix A, the asymptotic expansion of the corresponding integrals in the limit $\varepsilon \to 0$ gives

$$\mathcal{H}_f\varphi = \Delta\varphi - \varphi\big(e^U \Delta e^{-U}\big). \tag{22}$$

Similarly, the inifinitesimal operator of the backward operator is given by

$$\mathcal{H}_b\psi = \lim_{\varepsilon\to 0} \frac{T_b - I}{\varepsilon}\psi = \Delta\psi - 2\nabla\psi\cdot\nabla U. \tag{23}$$

As expected, $\psi_0 = 1$ is the eigenfunction with $\lambda_0 = 0$ of the backward infinitesimal operator, while $\varphi_0 = e^{-U}$ is the eigenfunction of the forward one. Thus, if N is large enough and ε is small enough, the structure of the right eigenvectors of the finite matrix M is similar to those of the eigenfunctions of the infinitesimal operators $\mathcal{H}_b$.

A few important remarks are due at this point. First, note that the backward operator (23) has the same functional form as the backward FPE (5), but with a potential $2U(\mathbf{x})$ instead of $U(\mathbf{x})$. The forward operator (22) has a different functional form from the forward FPE (3) corresponding to the stochastic differential equation (2). This should come as no surprise, since (22) is the differential operator of an isotropic diffusion process on a space with non-uniform probability measure $\mu(\mathbf{x})$, which is obviously different from the standard anisotropic diffusion in a space with a uniform measure, as described by the SDE (2) [21].

Interestingly, however, the form of the forward operator is the same as the Schrödinger operator of quantum physics [25], e.g.,

$$\mathcal{H}\varphi = \Delta\varphi - \varphi V(\mathbf{x}), \tag{24}$$

where in our case the potential $V(\mathbf{x})$ has the following specific form:

$$V(\mathbf{x}) = \left\|\nabla U(\mathbf{x})\right\|^2 - \Delta U(\mathbf{x}). \tag{25}$$

Therefore, in the limit $N \to \infty, \varepsilon \to 0$, the left eigenvectors of the Markov matrix M converge to the eigenfunctions of the Schrödinger operator (24) with a potential (25). The properties of the first few of these eigenfunctions have been extensively studied for simple potentials $V(\mathbf{x})$ [25].

In order to see why the forward operator $\mathcal{H}_f$ also corresponds to a potential $2U(\mathbf{x})$ instead of $U(\mathbf{x})$, we recall the correspondence [18], between the Schrödinger equation with a supersymmetric potential of the specific form (25) and a diffusion process described by the Fokker–Planck equation (3). The transformation

$$p(\mathbf{x}, t) = \psi(\mathbf{x}, t)e^{-U(\mathbf{x})/2} \tag{26}$$

applied to the original FPE (3) yields the Schrödinger equation with imaginary time

$$-\frac{\partial\psi}{\partial t} = \Delta\psi - \psi\left(\frac{\|\nabla U\|^2}{4} - \frac{\Delta U}{2}\right). \tag{27}$$

Comparing (27) with (25), we conclude that the eigenvalues of the operator (22) are the same as those of a Fokker–Planck equation with a potential $2U(\mathbf{x})$. Therefore, in general, for data sampled from the SDE (2), there is no direct correspondence between the eigenvalues and eigenfunctions of the normalized graph Laplacian and those of the corresponding Fokker–Planck equation (3). However, when the original potential $U(\mathbf{x})$ has two metastable states separated by a large barrier, corresponding to two well separated clusters, so does $2U(\mathbf{x})$. Therefore, the first non-trivial eigenvalue is governed by the mean passage time between the two barriers and the first non-trivial eigenfunction gives a parametrization of the path between them (see also the analysis in the next section).

We note that in [26], Horn and Gottlieb suggested a clustering algorithm based on the Schrödinger operator (24), where they constructed an approximate eigenfunction $\psi(\mathbf{x}) = p_\varepsilon(\mathbf{x})$ as in our equation (11) and computed its corresponding potential $V(\mathbf{x})$ from Eq. (24). The clusters were then defined by the minima of the potential V. Employing a similar asymptotic analysis, one can show that in the appropriate limit, the computed potential V is given by (25). This asymptotic analysis and the connection between the quantum operator and a diffusion process thus provides further mathematical insight for the success of their method. Indeed, when U has a deep parabolic minima at a point $\mathbf{x}$, corresponding to a well-defined cluster, so does V.

4. Anisotropic diffusion maps

As shown in the previous section, the eigenvalues and eigenvectors of the normalized graph Laplacian operator correspond to those of a Fokker–Planck operator with a potential $2U(\mathbf{x})$ instead of the single $U(\mathbf{x})$. In this section we present a different normalization that yields infinitesimal generators corresponding to the potential $U(\mathbf{x})$ without the additional factor of two.

In fact, following [7] we consider in more generality a whole family of different normalizations and their corresponding diffusions, and we show that, in addition to containing the graph Laplacian normalization of the previous

section, this collection of diffusions includes at least two other Laplacians of interest: the Laplace–Beltrami operator, which captures the Riemannian geometry of the data set, and the backward Fokker–Planck operator of Eq. (5).

Instead of applying the graph Laplacian normalization to the isotropic kernel $k_\varepsilon(\mathbf{x},\mathbf{y})$, we first appropriately adapt the kernel into an anisotropic one to obtain a new weighted graph, to which we apply the random walk graph Laplacian normalization. More precisely, we proceed as follows: start with a Gaussian kernel $k_\varepsilon(\mathbf{x},\mathbf{y})$ and let $\alpha > 0$ be a parameter indexing our family of diffusions. Define an estimate for the local density as

$$p_\varepsilon(\mathbf{x}) = \int k_\varepsilon(\mathbf{x},\mathbf{y}) \Pr\{\mathbf{y}\}\, d\mathbf{y}$$

and consider the family of kernels

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

We now apply the graph Laplacian normalization by computing the normalization factor

$$d_\varepsilon^{(\alpha)}(\mathbf{y}) = \int k_\varepsilon^{(\alpha)}(\mathbf{x},\mathbf{y}) \Pr\{\mathbf{x}\}\, d\mathbf{x}$$

and forming a forward transition probability kernel

$$M_f^{(\alpha)}(\mathbf{x}\mid\mathbf{y}) = \Pr\big\{\mathbf{x}(t+\varepsilon) = \mathbf{x} \mid \mathbf{x}(t) = \mathbf{y}\big\} = \frac{k_\varepsilon^{(\alpha)}(\mathbf{x},\mathbf{y})}{d_\varepsilon^{(\alpha)}(\mathbf{y})}.$$

Similar to the analysis of Section 3.2, we can construct the corresponding forward, symmetric and backward diffusion kernels. It can be shown (see Appendix A) that the forward and backward infinitesimal generators of this diffusion are

$$\mathcal{H}_b^{(\alpha)}\psi = \Delta\psi - 2(1-\alpha)\nabla\phi\cdot\nabla U, \tag{28}$$

$$\mathcal{H}_f^{(\alpha)}\varphi = \Delta\varphi - 2\alpha\nabla\varphi\cdot\nabla U + (2\alpha-1)\varphi\big((\nabla U)^2 - \Delta U\big). \tag{29}$$

We mention three interesting cases:

- For $\alpha = 0$, this construction yields the classical normalized graph Laplacian with the infinitesimal operator given by Eq. (23)

  $$\mathcal{H}_b\psi = \Delta\psi - 2\nabla U\cdot\nabla\psi.$$

- For $\alpha = 1$, the backward generator gives the Laplace–Beltrami operator:

  $$\mathcal{H}_b\psi = \Delta\psi. \tag{30}$$

  In other words, this diffusion captures only the geometry of the data (e.g., the domain Ω), with the density e^{-U} playing absolutely no role. Therefore, this normalization separates the geometry of the underlying manifold from the statistics on it.
- For $\alpha = \frac{1}{2}$, the infinitesimal operator of the forward and backward operators coincide and are given by

  $$\mathcal{H}_f\varphi = \mathcal{H}_b\varphi = \Delta\varphi - \nabla\varphi\cdot\nabla U \tag{31}$$

  which is exactly the backward FPE (5), with a potential $U(\mathbf{x})$.

Therefore, the last case with $\alpha = 1/2$ provides a consistent method to approximate the eigenvalues and eigenfunctions corresponding to the stochastic differential equation (2). This is done by constructing a graph Laplacian with an appropriately anisotropic weighted graph.

As discussed above and in more detail in [7,13,30], in the presence of a spectral gap, the Euclidean distance between any two points after the diffusion map embedding into $\mathbb{R}^k$ is almost equal to their diffusion distance on the original dataset. Thus, for dynamical systems with only a few slow time scales and many fast time scales, only a small number of diffusion map coordinates need be retained for the coarse grained representation of the data at medium to long times, at which the fast coordinates have equilibrated. Therefore, the diffusion map can be considered as an empirical method to perform data-driven or equation-free homogenization. In particular, since this observation yields

a computational method for the approximation of the top eigenfunctions and eigenvalues, this method can be applied toward the design of fast and efficient simulations that can be initialized on arbitrary points on the diffusion map. This application is described in a separate publication [30].

5. Examples

In this section we present the potential strength of the diffusion map method by analyzing, both analytically and numerically a few toy examples with simple potentials $U(\mathbf{x})$. More complicated high dimensional examples of stochastic dynamical systems are analyzed in [30], while other applications such as the analysis of images for which we typically have no underlying probability model appear in [7]. An example where the density plays no role but the geometry Ω defines the structure of the eigenvalues and eigenvectors is described in [23].

5.1. Parabolic potential in 1-D

We start with the simplest case of a parabolic potential in one dimension, which in the context of the SDE (2), corresponds to the well known Ornstein–Uhlenbeck process. We thus consider a potential $U(x) = x^2/2\tau$, with a corresponding normalized density $p = e^{-U}/\sqrt{2\pi\tau}$.

The normalization factor p_ε can be computed explicitly

$$p_\varepsilon(y) = \int_{-\infty}^{\infty} \frac{e^{-(x-y)^2/2\varepsilon}}{\sqrt{2\pi\varepsilon}} \frac{e^{-x^2/2\tau}}{\sqrt{2\pi\tau}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi(\tau+\varepsilon)}} e^{-y^2/2(\tau+\varepsilon)}$$

where, for convenience, we multiplied the kernel $k_\varepsilon(x, y)$ by a normalization factor $1/\sqrt{2\pi\varepsilon}$. Therefore, the eigenvalue/eigenfunction problem for the symmetric operator T_s with a finite ε reads

$$T_s\varphi = \int_{-\infty}^{\infty} \frac{\exp\left(-\frac{(x-y)^2}{2\varepsilon}\right)}{\sqrt{2\pi\varepsilon}} \exp\left(\frac{x^2+y^2}{4(\varepsilon+\tau)}\right) \exp\left(-\frac{y^2}{2\tau}\right) \sqrt{\frac{\tau+\varepsilon}{\tau}}\varphi(y)\,\mathrm{d}y = \lambda\varphi(x).$$

The first eigenfunction, with eigenvalue $\lambda_0 = 1$ is given by

$$\varphi_0(x) = C\sqrt{p_\varepsilon(x)} = C\exp\left(-\frac{x^2}{4(\varepsilon+\tau)}\right).$$

The second eigenfunction with eigenvalue $\lambda_1 = \tau/(\tau+\varepsilon) < 1$ is, up to normalization,

$$\varphi_1(x) = x\exp\left(-\frac{x^2}{4(\varepsilon+\tau)}\right).$$

In general, the sequence of eigenfunctions and eigenvalues is characterized by the following lemma.

Lemma. *The eigenvalues of the operator T_s are $\lambda_k = (\tau/(\tau+\varepsilon))^k$, with the corresponding eigenvectors given by*

$$\varphi_k(x) = p_k(x)\exp\left(-\frac{x^2}{4(\tau+\varepsilon)}\right), \tag{32}$$

where p_k is a polynomial of degree k (even or odd depending on k).

In the limit $\varepsilon \to 0$ we obtain the eigenfunctions of the corresponding infinitesimal generator. For the specific potential $U(x) = x^2/2\tau$, the eigenfunction problem for the backward generator reads

$$\psi_{xx} - 2\frac{x}{\tau}\psi_x = -\lambda\psi \tag{33}$$

and its solutions are the well-known Hermite polynomials. Due to the relation between this operator and the Schrödinger operator, these are also the eigenfunctions of the quantum harmonic oscillator (after multiplication by the appropriate Gaussian) [25].

Note that plotting the second vs the first backward eigenfunctions (with the convention that the zeroth eigenfunction is the constant one, which we typically ignore) is the same as plotting $x^2 + 1$ vs x, e.g., a parabola. Therefore, we expect that for a large enough and yet finite data-set sampled from this potential, the plot of the corresponding discrete eigenfunctions should lay on a parabolic curve (see next section for a numerical example).

5.2. Multi-dimensional parabolic potential

We now consider a harmonic potential in n dimensions of the form

$$U(\mathbf{x}) = \sum_j \frac{x_j^2}{2\tau_j} \tag{34}$$

where, in addition, we assume $\tau_1 \gg \tau_2 > \tau_3 > \cdots > \tau_n$, so that x_1 is a slow variable in the context of the SDE (2).

We note that for this specific potential, the probability density has a separable structure, $p(\mathbf{x}) = p_1(x_1) \cdots p_n(x_n)$, and so does the kernel $k_\varepsilon(\mathbf{x}, \mathbf{y})$, and consequently, also the symmetric kernel $M_s(\mathbf{x}, \mathbf{y})$. Therefore, there is an outer-product structure to the eigenvalues and eigenfunctions of the integral operators T_f, T_s, T_b. For example, in two dimensions the eigenfunctions and eigenvalues are given by

$$\varphi_{i,j}(x_1, x_2) = \varphi_{1,i}(x_1)\varphi_{2,j}(x_2) \quad \text{and} \quad \lambda_{i,j} = \mu_1^i \mu_2^j, \tag{35}$$

where $\mu_1 = \tau_1/(\tau_1 + \varepsilon)$ and $\mu_2 = \tau_2/(\tau_2 + \varepsilon)$. Since by assumption $\tau_1 \gg \tau_2$, then upon ordering of the eigenfunctions by decreasing eigenvalue, the first non-trivial eigenfunctions are $\varphi_{1,0}, \varphi_{2,0}, \ldots$, which depend only on the slow variable x_1. Note that indeed, regardless of the value of ε, as long as $\tau_2 > 2\tau_1$ we have that $\lambda_1^2 > \lambda_2$. Therefore, in this example the first few coordinates of the diffusion map give a (redundant) parametrization of the slow variable x_1 in the system.

In Fig. 1 we present numerical results corresponding to a 2-dimensional potential with $\tau_1 = 1, \tau_2 = 1/25$. In the upper left some 3500 points sampled from the distribution $p = e^{-U}$ are shown. In the lower right and left panels, the first two non-trivial backward eigenfunctions ψ_1 and ψ_2 are plotted vs the slow variable x_1. Note that except at the edges, where the statistical sampling is poor, the first eigenfunction is linear in x_1 while the second one is quadratic in x_1. In the upper right panel we plot ψ_2 vs ψ_1 and note that they indeed lie on a parabolic curve, as predicted by the analysis of the previous section.

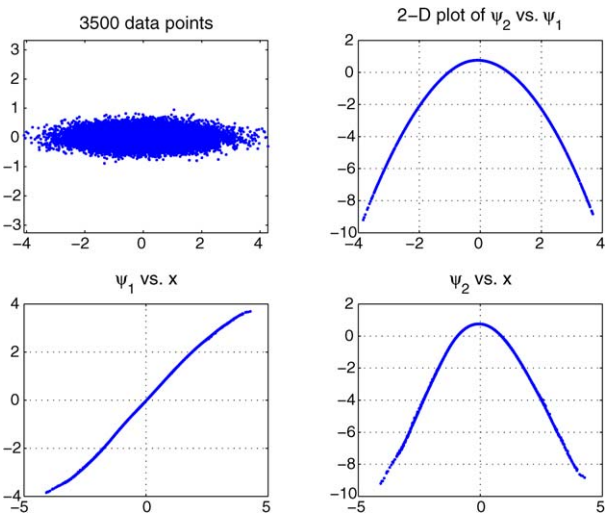

Fig. 1. The anisotropic diffusion map on a harmonic potential in 2-D.

5.3. A potential with two minima

We now consider a double well potential $U(x)$ with two minima, one at x_L and one at x_R (see Fig. 2). For simplicity, we analyze the case of a symmetric potential around $(x_L + x_R)/2$ with $U(x_L) = U(x_R) = 0$. In the context of data clustering this can be viewed as approximately a mixture of two Gaussian clouds, while in the context of stochastic dynamical systems this potential defines two meta-stable states.

We first consider an approximation to the quantity $p_\varepsilon(x)$, given by Eq. (13). For x near x_L, $U(x) \approx (x - x_L)^2/\tau_L$, while for x near x_R, $U(x) \approx (x - x_R)^2/\tau_R$. Therefore,

$$e^{-U(y)} \approx e^{-(y-x_L)^2/2\tau_L} + e^{-(y-x_R)^2/2\tau_R} \tag{36}$$

and

$$p_\varepsilon(x) \approx \frac{1}{\sqrt{2}}\left(\frac{\sqrt{\tau_L}}{\sqrt{\tau_L+\varepsilon}} e^{-(x-x_L)^2/2(\tau_L+\varepsilon)} + \frac{\sqrt{\tau_R}}{\sqrt{\tau_R+\varepsilon}} e^{-(x-x_R)^2/2(\tau_R+\varepsilon)}\right) = \frac{1}{\sqrt{2}}\big[\varphi_L(x) + \varphi_R(x)\big], \tag{37}$$

where φ_L and φ_R are the first forward eigenfunctions corresponding to a single well potential centered at x_L or at x_R, respectively. As is well known both in the theory of quantum physics and in the theory of the Fokker–Planck equation, an approximate expression for the next eigenfunction is

$$\varphi_1(x) = \frac{1}{\sqrt{2}}\big[\varphi_L(x) - \varphi_R(x)\big].$$

Therefore, the first non-trivial eigenfunction of the backward operator is given by

$$\psi_1(x) = \frac{\varphi_L(x) - \varphi_R(x)}{\varphi_L(x) + \varphi_R(x)}.$$

This eigenfunction is roughly $+1$ in one well and -1 in the other well, with a sharp transition between the two values near the barrier between the two wells. Therefore, this eigenfunction is indeed suited for clustering. Moreover, in the context of a mixture of two Gaussian clouds, clustering according to the sign of $\psi_1(x)$ is asymptotically equivalent to the optimal Bayes classifier.

Example. Consider the following potential in two dimensions:

$$U(x, y) = \frac{1}{4}x^4 - \frac{25}{12}x^3 + \frac{9}{2}x^2 + 25\frac{y^2}{2}. \tag{38}$$

In the x direction, this potential has a double well shape with two minima, one at $x = 0$ and one at $x = 4$, separated by a potential barrier with a maximum at $x = 2.25$.

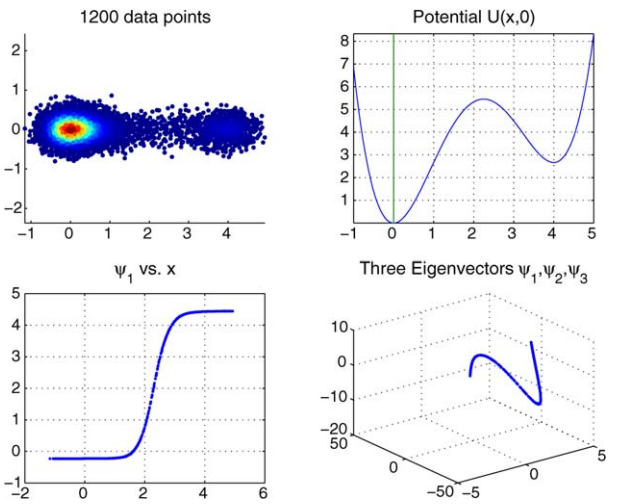

Fig. 2. Numerical results for a double well potential in 2-D.

In Fig. 2 we show some numerical results of the diffusion map on some 1200 points sub-sampled from a stochastic simulation with this potential which generated about 40,000 points. On the upper right panel we plotted the potential $U(x,0)$, showing the two wells. In the upper left, a scatter plot of all the points, color coded according to the value of the local estimated density p_ε (with $\varepsilon = 0.25$), is shown, where the two clusters are easily observed. In the lower left panel, the first non-trivial eigenfunction is plotted vs the first coordinate x. Note that even though there is quite a bit of variation in the y-variable inside each of the wells, the first eigenfunction ψ_1 is essentially a function of only x, regardless of the value of y. In the lower right we plot the first three backward eigenfunctions. Note that they all lie on a curve, indicating that the long time asymptotics are governed by a *single* time scale, the passage time between the two wells, and not by the local fluctuations inside them.

5.4. Potential with three wells

We now consider the following two dimensional potential energy with three wells:

$$U(x,y) = 3\beta e^{-x^2}\left[e^{-(y-1/3)^2} - e^{-(y-5/3)^2}\right] - 5\beta e^{-y^2}\left[e^{-(x-1)^2} + e^{-(x+1)^2}\right], \tag{39}$$

where $\beta = 1/kT$ is a thermal factor. This potential has two deep wells near $(-1,0)$ and $(1,0)$ and a shallower well near $(0,5/3)$, which we denote as the points L, R, C, respectively, The transitions between the wells of this potential have been analyzed in many works [27]. In Fig. 3 we plotted on the left the results of 1400 points sub-sampled from a total of 80,000 points randomly generated from this potential confined to the region $[-2.5,2.5]^2 \subset \mathbb{R}^2$ at temperature $\beta = 2$, color-coded by their local density. On the right we plotted the first two diffusion map coordinates $\psi_1(\mathbf{x}), \psi_2(\mathbf{x})$. Notice that in the diffusion map coordinates, the majority of the sampled points get mapped into a triangle where each vertex corresponds to one of the points L, R, C. This figure shows that there are two possible pathways to go from L to R. A direct (short) way and an indirect longer way, that passes through the shallow well centered at C.

5.5. Iris data set

We conclude this section with a diffusion map analysis of one of the most popular multivariate datasets in pattern recognition, the iris data set. This set contains 3 distinct classes of samples in four dimensions, with 50 samples in each class. In Fig. 4 we see on the left the result of the three-dimensional diffusion map on this dataset. In the diffusion map coordinates, all 50 points of class 1 (blue; for interpretation of the references to color see web version of this article) are shrunk into a single point in the diffusion map space and can thus be easily distinguished from classes two and three (red and green; for interpretation of the references to color see web version of this article). In the right plot we see the results of re-running the diffusion map on the 100 remaining red and green samples. The 2-D plot of the

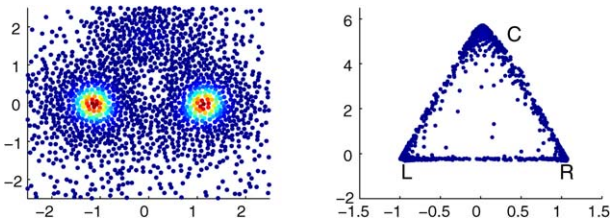

Fig. 3. Numerical results for a triple well potential in 2-D.

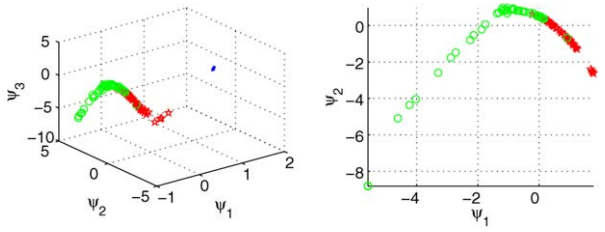

Fig. 4. Diffusion map for the iris data set.

first two diffusion maps coordinates shows that there is no perfect separation between these two classes. However, clustering according to the sign of $\psi_1(\mathbf{x})$ gives misclassifications rates similar to those of other methods, of the order of 6–8 samples depending on the value chosen for the kernel width ε.

6. Summary and discussion

In this paper, we introduced a mathematical framework for the analysis of diffusion maps, via their corresponding infinitesimal generators. Our results show that diffusion maps are a natural method for the analysis of the geometry and probability distribution of empirical data sets. The identification of the eigenvectors of the Markov chain as discrete approximations to the corresponding differential operators provides a mathematical justification for their use as a dimensional reduction tool and gives an alternative explanation for their empirical success in various data analysis applications, such as spectral clustering and approximations of optimal normalized cuts on discrete graphs [23].

We generalized the standard construction of the normalized graph Laplacian to a one-parameter family of graph Laplacians that provides a low-dimensional description of the data combining the geometry of the set with the probability distribution of the data points. The choice of the diffusion map depends on the task at hand. If, for example, data points are known to approximately lie on a manifold, and one is solely interested in recovering the geometry of this set, then an appropriate normalization of a Gaussian kernel allows to approximate the Laplace–Beltrami operator, regardless of the density of the data points. This construction achieves a complete separation of the underlying geometry, represented by the knowledge of the Laplace operator, from the statistics of the points. This is important in situations where the density is meaningless and yet points on the manifold are not sampled uniformly on it. In a different scenario, if the data points are known to be sampled from the equilibrium distribution of a Fokker–Planck equation, the long-time dynamics of the density of points can be recovered from an appropriately normalized random walk process. In this case, there is a subtle interaction between the distribution of the points and the geometry of the data set, and one must correctly account for both.

While in this paper we analyzed only Gaussian kernels, our asymptotic results are valid for general kernels, with the appropriate modification that take into account the mean and covariance matrix of the kernel. Note, however, that although asymptotically in the limit $N \to \infty$ and $\varepsilon \to 0$, the choice of the isotropic kernel is unimportant, for a finite data set the choice of both ε and the kernel can be crucial for the success of the method.

Finally, in the context of dynamical systems, we showed that diffusion maps with the appropriate normalization constitute a powerful tool for the analysis of systems exhibiting different time scales. In particular, as shown in the different examples, these time scales can be separated and the long time dynamics can be characterized by the top eigenfunctions of the diffusion operator. Last, our analysis paves the way for fast simulations of physical systems by allowing larger integration steps along slow variable directions. The exact details required for the design of fast and efficient simulations based on diffusion maps will be described in a separate publication [30].

Acknowledgment

The authors thank the referee for helpful suggestions and for pointing out Ref. [26].

Appendix A. Infinitesimal operators for a family of graph Laplacians

In this appendix, we present the calculation of the infinitesimal generators for the different diffusion maps characterized by a parameter α.

Suppose that the data set X consists of a Riemannian manifold with a density $p(\mathbf{x}) = e^{-U(\mathbf{x})}$ and let $k_\varepsilon(\mathbf{x}, \mathbf{y})$ be a Gaussian kernel. It was shown in [7] that if k_ε is scaled appropriately, then for any function ϕ on X,

$$\int_X k_\varepsilon(\mathbf{x}, \mathbf{y})\phi(\mathbf{y})\,\mathrm{d}y = \phi(\mathbf{x}) + \varepsilon\big(\Delta\phi(x) + q(\mathbf{x})\phi(\mathbf{x})\big) + O\big(\varepsilon^{\frac{3}{2}}\big),$$

where q is a function that depends on the Riemannian geometry of the manifold and its embedding in $\mathbb{R}^n$. Using the notations introduced in Section 4, it is easy to verify that

$$p_\varepsilon(\mathbf{x}) = p(\mathbf{x}) + \varepsilon\big(\Delta p(\mathbf{x}) + q(\mathbf{x})p(\mathbf{x})\big) + O\big(\varepsilon^{3/2}\big)$$

and consequently,

$$p_\varepsilon^{-\alpha} = p^{-\alpha}\left(1 - \alpha\varepsilon\left(\frac{\Delta p}{p} + q\right)\right)\left(1 + O\left(\varepsilon^{3/2}\right)\right).$$

Let

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

Then, the normalization factor $d_\varepsilon^{(\alpha)}$ is given by

$$d_\varepsilon^{(\alpha)}(\mathbf{x}) = \int k_\varepsilon^{(\alpha)}(\mathbf{x},\mathbf{y})\,p(\mathbf{y})\,d\mathbf{y} = p_\varepsilon^{-\alpha}(\mathbf{x})\,p^{1-\alpha}(\mathbf{x})\left[1 + \varepsilon\left((1-\alpha)q - \alpha\frac{\Delta p}{p} + \frac{\Delta p^{1-\alpha}}{p^{1-\alpha}(\mathbf{x})}\right)\right].$$

Therefore, the asymptotic expansion of the backward operator gives

$$T_b^{(\alpha)}\phi = \int_X \frac{k_\varepsilon^{(\alpha)}(\mathbf{x},\mathbf{y})}{d_\varepsilon^{(\alpha)}(\mathbf{x})}\phi(\mathbf{y})\,p(\mathbf{y})\,d\mathbf{y} = \phi(\mathbf{x}) + \varepsilon\left(\frac{\Delta(\phi p^{1-\alpha})}{p^{1-\alpha}} - \phi\frac{\Delta p^{1-\alpha}}{p^{1-\alpha}}\right)$$

and its infinitesimal generator is

$$\mathcal{H}_b\phi = \lim_{\varepsilon\to 0}\frac{T_b - I}{\varepsilon}\phi = \frac{\Delta(\phi p^{1-\alpha})}{p^{1-\alpha}} - \frac{\Delta(p^{1-\alpha})}{p^{1-\alpha}}\phi.$$

Inserting the expression $p = e^{-U}$ into the last equation gives

$$\mathcal{H}_b\phi = \Delta\phi - 2(1-\alpha)\nabla\phi\cdot\nabla U.$$

Similarly, the form of the forward infinitesimal operator is

$$\mathcal{H}_f\psi = \Delta\psi - 2\alpha\nabla\psi\cdot\nabla U + (2\alpha - 1)\psi(\nabla U\cdot\nabla U - \Delta U).$$

References

[1] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, second ed., Springer-Verlag, New York, 2001.
[2] 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, Part I: Diffusion maps, Proc. Natl. Acad. Sci. 102 (21) (2005) 7426–7431.
[3] D. Givon, R. Kupferman, A. Stuart, Extracting macroscopic dynamics: Model problems and algorithms, Nonlinearity 17 (2004) R55–R127.
[4] W. Huisinga, C. Best, R. Roitzsch, Ch. Schütte, F. Cordes, From simulation data to conformational ensembles: Structure and dynamics based methods, J. Comput. Chem. 20 (1999) 1760–1774.
[5] W. Huisinga, Ch. Schütte, A.M. Stuart, Extracting macroscopic stochastic dynamics: Model problems, Comm. Pure Appl. Math. 56 (2003) 234–269.
[6] T. Faradjian, R. Elber, Computing time scales from reaction coordinates by milestoning, J. Chem. Phys. 120 (2004) 10880–10889.
[7] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 6–31.
[8] F.R.K. Chung, Spectral Graph Theory, Reg. Conf. Ser. Math., vol. 92, Amer. Math. Soc., Providence, RI, 1997.
[9] Y. Weiss, Segmentation using eigenvectors: A unifying view, in: Proc. IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 975–982.
[10] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, Adv. Neural Inform. Process. Syst. 14 (2002).
[11] J. Shi, J. Malik, Normalized cuts and image segmentation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 731–737.
[12] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[13] M. Saerens, F. Fouss, L. Yen, P. Dupont, The principal components analysis of a graph and its relationships to spectral clustering, in: Proc. 15th European Conference on Machine Learning, ECML, 2004, in: Lecture Notes in Artificial Intelligence, vol. 3201, Springer-Verlag, Berlin, 2004, pp. 371–383.
[14] M. Belkin, P. Niyogi, Toward a theoretical foundation for Laplacian based manifold methods, in: 18th Conference on Learning Theory, 2005.
[15] M. Meila, J. Shi, A random walks view of spectral segmentation, AI and Statistics, 2001.
[16] L. Yen, D. Vanvyve, F. Wouters, F. Fouss, M. Verleysen, M. Saerens, Clustering using a random-walk based distance measure, in: European Symposium on Artificial Neural Networks, 2005, pp. 317–324.
[17] N. Tishby, N. Slonim, Data clustering by Markovian relaxation and the information bottleneck method, Adv. Neural Inform. Process. Syst. 13 (2001).
[18] M. Bernstein, L.S. Brown, Supersymmetry and the bistable Fokker–Planck equation, Phys. Rev. Lett. 52 (1984) 1933–1935.

[19] I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, C. Theodoropoulos, Equation free multiscale computation: Enabling microscopic simulators to perform system-level tasks, Commun. Math. Sci. 1 (4) (2003) 715.
[20] Z. Schuss, Theory and Applications of Stochastic Differential Equations, Wiley, New York, 1980.
[21] C.W. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, third ed., Springer-Verlag, New York, 2004.
[22] H. Risken, The Fokker–Planck Equation: Methods of Solution and Applications, Springer-Verlag, Berlin/New York, 1989.
[23] B. Nadler, S. Lafon, I.G. Kevrekidis, R.R. Coifman, Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators, Adv. Neural Inform. Process. Syst. 18 (2005).
[24] J. Ham, D.D. Lee, S. Mika, B. Schölkopf, A kernel view of the dimensionality reduction of manifolds, Technical report TR-110, Max-Planck-Institut für biologische Kybernetik, Tübingen, 2003.
[25] J. Singh, Quantum Mechanics, Wiley, New York, 1997.
[26] D. Horn, A. Gottlieb, Algorithm for data clustering in pattern recognition problems based on quantum mechanics, Phys. Rev. Lett. 88 (1) (2002) 018702.
[27] S. Park, M.K. Sener, D. Lu, K. Schulten, Reaction paths based on mean first passage times, J. Chem. Phys. 119 (3) (2003) 1313–1319.
[28] U. von Luxburg, O. Bousquet, M. Belkin, Limits of spectral clustering, Adv. Neural Inform. Process. Syst. 17 (2004).
[29] M. Hein, J. Audibert, U. von Luxburg, From graphs to manifolds—Weak and strong pointwise consistency of graph Laplacians, in: 18th Conference on Learning Theory, 2005.
[30] R.R. Coifman, S. Lafon, M. Maggioni, I.G. Kevrekidis, B. Nadler, Diffusion maps, reaction coordinates and low dimensional representation of stochastic systems, in preparation.