Key Points in Coifman-Lafon

Here we will include key parts of the paper after the students are already trained in these notions.

Coifman and Lafon, "Diffusion Maps", Appl Comput Harmon Analysis, Vol 21 (2006) 5-30. (reprint)

and explain them for students enrolled in Geometry Research with Professors Lin and Sormani in Summer 2020.

This document is not yet complete.

ELSEVIER

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

We will use specific h and k in our work, some of this is more general.

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).

2.4. 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:

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

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

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

then, up to relative precision δ, we have

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

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

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

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.

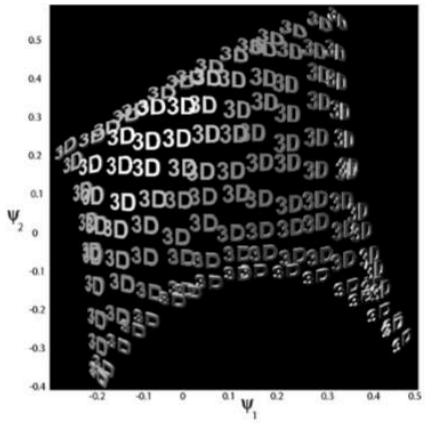

Fig. 2. The set of images reorganized by the two eigenvectors ψ_1 and ψ_2. We recover the natural organization dictated by the angles of rotation.

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.

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 $\mathcal{M}$:

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

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

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

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 $\mathcal{M}$.

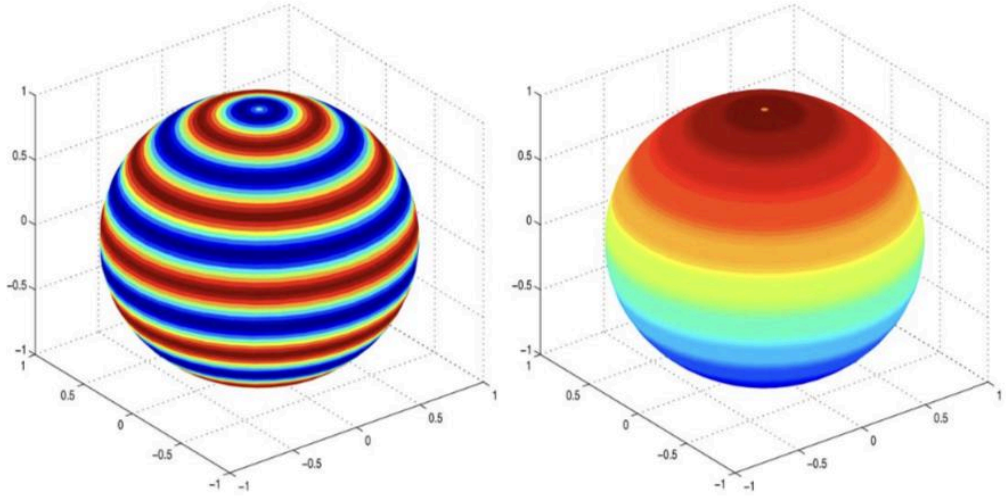

Fig. 5. Left: the original function $f(\varphi, \theta) = \sin(12\theta)$. Right: first nontrivial eigenfunction $\phi_1(\varphi, \theta) = \cos(\theta)$.

Delete reference to Markov chains below.

We will use: h(r)=e^r

Weighted laplacian

Heat kernel

Below introduce graph setting:

So now x and y are bunch of dots on the manifold

Need to define the graph Laplacian

alpha is a power not an upper index

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

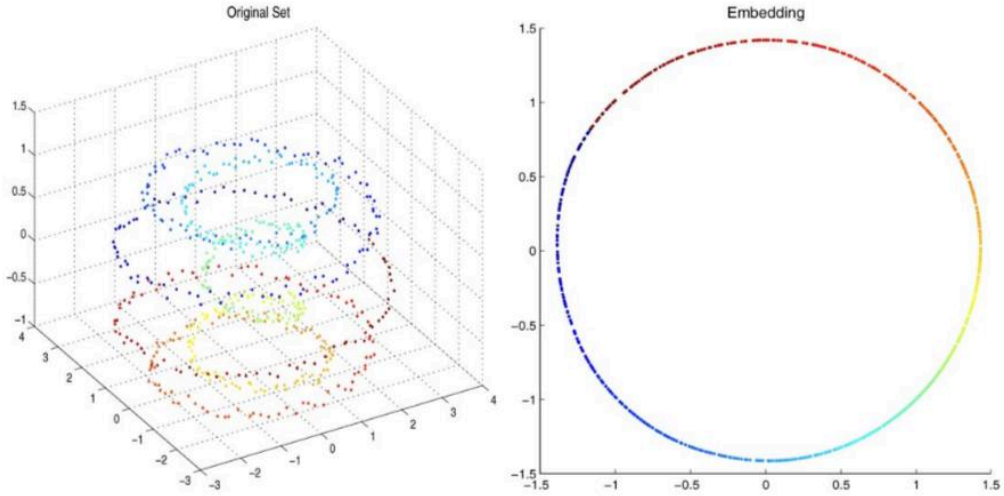

Fig. 6. Left: a noisy helix. Right: the curve is embedded as a perfect circle.