# Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators

**Boaz Nadler**[*] **Stéphane Lafon** **Ronald R. Coifman**
Department of Mathematics, Yale University, New Haven, CT 06520.
{boaz.nadler,stephane.lafon,ronald.coifman}@yale.edu

**Ioannis G. Kevrekidis**
Department of Chemical Engineering and Program in Applied Mathematics
Princeton University, Princeton, NJ 08544
yannis@princeton.edu

## Abstract

This paper presents a diffusion based probabilistic interpretation of spectral clustering and dimensionality reduction algorithms that use the eigenvectors of the normalized graph Laplacian. Given the pairwise adjacency matrix of all points, we define a diffusion distance between any two data points and show that the low dimensional representation of the data by the first few eigenvectors of the corresponding Markov matrix is optimal under a certain mean squared error criterion. Furthermore, assuming that data points are random samples from a density $p(\boldsymbol{x}) = e^{-U(\boldsymbol{x})}$ we identify these eigenvectors as discrete approximations of eigenfunctions of a Fokker-Planck operator in a potential $2U(\boldsymbol{x})$ with reflecting boundary conditions. Finally, applying known results regarding the eigenvalues and eigenfunctions of the continuous Fokker-Planck operator, we provide a mathematical justification for the success of spectral clustering and dimensional reduction algorithms based on these first few eigenvectors. This analysis elucidates, in terms of the characteristics of diffusion processes, many empirical findings regarding spectral clustering algorithms.

**Keywords:** Algorithms and architectures, learning theory.

## 1 Introduction

Clustering and low dimensional representation of high dimensional data are important problems in many diverse fields. In recent years various spectral methods to perform these tasks, based on the eigenvectors of adjacency matrices of graphs on the data have been developed, see for example [1]-[10] and references therein. In the simplest version, known as the normalized graph Laplacian, given $n$ data points $\{\boldsymbol{x}_i\}_{i=1}^n$ where each $\boldsymbol{x}_i \in \mathbb{R}^p$, we define a pairwise similarity matrix between points, for example using a Gaussian kernel

[*]Corresponding author. Currently at Weizmann Institute of Science, Rehovot, Israel. http://www.wisdom.weizmann.ac.il/∼nadler

with width $\varepsilon$,

$$L_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\varepsilon}\right) \tag{1}$$

and a diagonal normalization matrix $D_{i,i} = \sum_j L_{i,j}$. Many works propose to use the first few eigenvectors of the normalized eigenvalue problem $L\phi = \lambda D\phi$, or equivalently of the matrix $M = D^{-1}L$, either as a low dimensional representation of data or as good coordinates for clustering purposes. Although eq. (1) is based on a Gaussian kernel, other kernels are possible. While for actual datasets the choice of a kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is crucial, it does not qualitatively change our asymptotic analysis [11].

The use of the first few eigenvectors of $M$ as good coordinates is typically justified with heuristic arguments or as a relaxation of a discrete clustering problem [3]. In [4, 5] Belkin and Niyogi showed that when data is uniformly sampled from a low dimensional manifold of $\mathbb{R}^p$ the first few eigenvectors of $M$ are discrete approximations of the eigenfunctions of the Laplace-Beltrami operator on the manifold, thus providing a mathematical justification for their use in this case. A different theoretical analysis of the eigenvectors of the matrix $M$, based on the fact that $M$ is a stochastic matrix representing a random walk on the graph was described by Meilă and Shi [12], who considered the case of piecewise constant eigenvectors for specific lumpable matrix structures. Additional notable works that considered the random walk aspects of spectral clustering are [8, 13], where the authors suggest clustering based on the average commute time between points, and [14] which considered the relaxation process of this random walk.

In this paper we provide a unified probabilistic framework which combines these results and extends them in two different directions. First, in section 2 we define a distance function between any two points based on the random walk on the graph, which we naturally denote the *diffusion distance*. We then show that the low dimensional description of the data by the first few eigenvectors, denoted as the *diffusion map*, is optimal under a mean squared error criterion based on this distance. In section 3 we consider a statistical model, in which data points are iid random samples from a probability density $p(\boldsymbol{x})$ in a smooth bounded domain $\Omega \subset \mathbb{R}^p$ and analyze the asymptotics of the eigenvectors as the number of data points tends to infinity. This analysis shows that the eigenvectors of the finite matrix $M$ are discrete approximations of the eigenfunctions of a Fokker-Planck (FP) operator with reflecting boundary conditions. This observation, coupled with known results regarding the eigenvalues and eigenfunctions of the FP operator provide new insights into the properties of these eigenvectors and on the performance of spectral clustering algorithms, as described in section 4.

## 2 Diffusion Distances and Diffusion Maps

The starting point of our analysis, as also noted in other works, is the observation that the matrix $M$ is adjoint to a symmetric matrix

$$M_s = D^{1/2} M D^{-1/2}. \tag{2}$$

Thus, $M$ and $M_s$ share the same eigenvalues. Moreover, since $M_s$ is symmetric it is diagonalizable and has a set of $n$ real eigenvalues $\{\lambda_j\}_{j=0}^{n-1}$ whose corresponding eigenvectors $\{\boldsymbol{v}_j\}$ form an orthonormal basis of $\mathbb{R}^n$. The left and right eigenvectors of $M$, denoted $\phi_j$ and $\psi_j$ are related to those of $M_s$ according to

$$\phi_j = \boldsymbol{v}_j D^{1/2}, \qquad \psi_j = \boldsymbol{v}_j D^{-1/2} \tag{3}$$

Since the eigenvectors $\boldsymbol{v}_j$ are orthonormal under the standard dot product in $\mathbb{R}^n$, it follows that the vectors $\phi_j$ and $\psi_k$ are bi-orthonormal

$$\langle \phi_i, \psi_j \rangle = \delta_{i,j} \tag{4}$$

where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ is the standard dot product between two vectors in $\mathbb{R}^n$. We now utilize the fact that by construction $M$ is a stochastic matrix with all row sums equal to one, and can thus be interpreted as defining a random walk on the graph. Under this view, $M_{i,j}$ denotes the transition probability from the point $\boldsymbol{x}_i$ to the point $\boldsymbol{x}_j$ in one time step. Furthermore, based on the similarity of the Gaussian kernel (1) to the fundamental solution of the heat equation, we define our time step as $\Delta t = \varepsilon$. Therefore,

$$\Pr\{\boldsymbol{x}(t+\varepsilon) = \boldsymbol{x}_j \,|\, \boldsymbol{x}(t) = \boldsymbol{x}_i\} = M_{i,j} \tag{5}$$

Note that $\varepsilon$ has therefore a *dual* interpretation in this framework. The first is that $\varepsilon$ is the (squared) radius of the neighborhood used to infer local geometric and density information for the construction of the adjacency matrix, while the second is that $\varepsilon$ is the discrete time step at which the random walk jumps from point to point.

We denote by $p(t, \boldsymbol{y}|\boldsymbol{x})$ the probability distribution of a random walk landing at location $\boldsymbol{y}$ at time $t$, given a starting location $\boldsymbol{x}$ at time $t = 0$. For $t = k\,\varepsilon$, $p(t, \boldsymbol{y}|\boldsymbol{x}_i) = \boldsymbol{e}_i M^k$, where $\boldsymbol{e}_i$ is a row vector of zeros with a single one at the $i$-th coordinate. For $\varepsilon$ large enough, all points in the graph are connected so that $M$ has a unique eigenvalue equal to 1. The other eigenvalues form a non-increasing sequence of non-negative numbers: $\lambda_0 = 1 > \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{n-1} \geq 0$. Then, regardless of the initial starting point $\boldsymbol{x}$,

$$\lim_{t \to \infty} p(t, \boldsymbol{y}|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) \tag{6}$$

where $\phi_0$ is the left eigenvector of $M$ with eigenvalue $\lambda_0 = 1$, explicitly given by

$$\phi_0(\boldsymbol{x}_i) = \frac{D_{i,i}}{\sum\limits_j D_{j,j}} \tag{7}$$

This eigenvector also has a dual interpretation. The first is that $\phi_0$ is the stationary probability distribution on the graph, while the second is that $\phi_0(\boldsymbol{x})$ is a density estimate at the point $\boldsymbol{x}$. Note that for a general shift invariant kernel $K(\boldsymbol{x} - \boldsymbol{y})$ and for the Gaussian kernel in particular, $\phi_0$ is simply the well known Parzen window density estimator.

For any finite time $t$, we decompose the probability distribution in the eigenbasis $\{\phi_j\}$

$$p(t, \boldsymbol{y}|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) + \sum_{j \geq 1} a_j(\boldsymbol{x}) \lambda_j^t \phi_j(\boldsymbol{y}) \tag{8}$$

where the coefficients $a_j$ depend on the initial location $\boldsymbol{x}$. Using the bi-orthonormality condition (4) gives $a_j(\boldsymbol{x}) = \psi_j(\boldsymbol{x})$, with $a_0(\boldsymbol{x}) = \psi_0(\boldsymbol{x}) = 1$ already implicit in (8).

Given the definition of the random walk on the graph it is only natural to quantify the similarity between any two points according to the evolution of their probability distributions. Specifically, we consider the following distance measure at time $t$,

$$\begin{aligned} D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) &= \|p(t, \boldsymbol{y}|\boldsymbol{x}_0) - p(t, \boldsymbol{y}|\boldsymbol{x}_1)\|_w^2 \\ &= \sum_{\boldsymbol{y}} (p(t, \boldsymbol{y}|\boldsymbol{x}_0) - p(t, \boldsymbol{y}|\boldsymbol{x}_1))^2 w(\boldsymbol{y}) \end{aligned} \tag{9}$$

with the specific choice $w(\boldsymbol{y}) = 1/\phi_0(\boldsymbol{y})$ for the weight function, which takes into account the (empirical) local density of the points.

Since this distance depends on the random walk on the graph, we quite naturally denote it as the *diffusion distance* at time $t$. We also denote the mapping between the original space and the first $k$ eigenvectors as the *diffusion map*

$$\Psi_t(\boldsymbol{x}) = \left(\lambda_1^t \psi_1(\boldsymbol{x}), \lambda_2^t \psi_2(\boldsymbol{x}), \ldots, \lambda_k^t \psi_k(\boldsymbol{x})\right) \tag{10}$$

The following theorem relates the diffusion distance and the diffusion map.

**Theorem:** The diffusion distance (9) is equal to Euclidean distance in the diffusion map space with all $(n-1)$ eigenvectors.

$$D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) = \sum_{j \geq 1} \lambda_j^{2t} \left(\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1)\right)^2 = \|\Psi_t(\boldsymbol{x}_0) - \Psi_t(\boldsymbol{x}_1)\|^2 \tag{11}$$

**Proof:** Combining (8) and (9) gives

$$D_t^2(\boldsymbol{x}_0, \boldsymbol{x}_1) = \sum_{\boldsymbol{y}} \Big( \sum_j \lambda_j^t (\psi_j(\boldsymbol{x}_0) - \psi_j(\boldsymbol{x}_1)) \phi_j(\boldsymbol{y}) \Big)^2 1/\phi_0(\boldsymbol{y}) \tag{12}$$

Expanding the brackets, exchanging the order of summation and using relations (3) and (4) between $\phi_j$ and $\psi_j$ yields the required result. Note that the weight factor $1/\phi_0$ is essential for the theorem to hold. $\square$.

This theorem provides a justification for using Euclidean distance in the diffusion map space for spectral clustering purposes. Therefore, geometry in diffusion space is meaningful and can be interpreted in terms of the Markov chain. In particular, as shown in [18], quantizing this diffusion space is equivalent to lumping the random walk. Moreover, since in many practical applications the spectrum of the matrix $M$ has a *spectral gap* with only a few eigenvalues close to one and all additional eigenvalues much smaller than one, the diffusion distance at a large enough time $t$ can be well approximated by only the first few $k$ eigenvectors $\psi_1(\boldsymbol{x}), \ldots, \psi_k(\boldsymbol{x})$, with a negligible error of the order of $O((\lambda_{k+1}/\lambda_k)^t)$. This observation provides a theoretical justification for dimensional reduction with these eigenvectors. In addition, the following theorem shows that this $k$-dimensional approximation is *optimal* under a certain mean squared error criterion.

**Theorem:** Out of all $k$-dimensional approximations of the form

$$\hat{p}(t, \boldsymbol{y}|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) + \sum_{j=1}^{k} a_j(t, \boldsymbol{x}) \boldsymbol{w}_j(\boldsymbol{y})$$

for the probability distribution at time $t$, the one that minimizes the mean squared error

$$\mathbb{E}_{\boldsymbol{x}} \{ \|p(t, \boldsymbol{y}|\boldsymbol{x}) - \hat{p}(t, \boldsymbol{y}|\boldsymbol{x})\|_w^2 \}$$

where averaging over initial points $\boldsymbol{x}$ is with respect to the stationary density $\phi_0(\boldsymbol{x})$, is given by $\boldsymbol{w}_j(\boldsymbol{y}) = \phi_j(\boldsymbol{y})$ and $a_j(t, \boldsymbol{x}) = \lambda_j^t \psi_j(\boldsymbol{x})$. Therefore, the optimal $k$-dimensional approximation is given by the truncated sum

$$\hat{p}(\boldsymbol{y}, t|\boldsymbol{x}) = \phi_0(\boldsymbol{y}) + \sum_{j=1}^{k} \lambda_j^t \psi_j(\boldsymbol{x}) \phi_j(\boldsymbol{y}) \tag{13}$$

**Proof:** The proof is a consequence of a weighted principal component analysis applied to the matrix $M$, taking into account the biorthogonality of the left and right eigenvectors.

We note that the first few eigenvectors are also optimal under other criteria, for example for data sampled from a manifold as in [4], or for multiclass spectral clustering [15].

## 3 The Asymptotics of the Diffusion Map

The analysis of the previous section provides a mathematical explanation for the success of the diffusion maps for dimensionality reduction and spectral clustering. However, it does not provide any information regarding the structure of the computed eigenvectors.

To this end, and similar to the framework of [16], we introduce a statistical model and assume that the data points $\{\boldsymbol{x}_i\}$ are i.i.d. random samples from a probability density $p(\boldsymbol{x})$

confined to a compact connected subset $\Omega \subset \mathbb{R}^p$ with smooth boundary $\partial\Omega$. Following the statistical physics notation, we write the density in Boltzmann form, $p(\boldsymbol{x}) = e^{-U(\boldsymbol{x})}$, where $U(\boldsymbol{x})$ is the (dimensionless) potential or energy of the configuration $\boldsymbol{x}$.

As shown in [11], in the limit $n \to \infty$ the random walk on the discrete graph converges to a random walk on the continuous space $\Omega$. Then, it is possible to define forward and backward operators $T_f$ and $T_b$ as follows,

$$T_f[\phi](\boldsymbol{x}) = \int_\Omega M(\boldsymbol{x}|\boldsymbol{y})\phi(\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y}, \quad T_b[\psi](\boldsymbol{x}) = \int_\Omega M(\boldsymbol{y}|\boldsymbol{x})\psi(\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y} \tag{14}$$

where $M(\boldsymbol{x}|\boldsymbol{y}) = \exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2\varepsilon)/D(\boldsymbol{y})$ is the transition probability from $\boldsymbol{y}$ to $\boldsymbol{x}$ in time $\varepsilon$, and $D(\boldsymbol{y}) = \int \exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2\varepsilon)p(\boldsymbol{x})d\boldsymbol{x}$.

The two operators $T_f$ and $T_b$ have probabilistic interpretations. If $\phi(\boldsymbol{x})$ is a probability distribution on the graph at time $t = 0$, then $T_f[\phi]$ is the probability distribution at time $t = \varepsilon$. Similarly, $T_b[\psi](\boldsymbol{x})$ is the mean of the function $\psi$ at time $t = \varepsilon$, for a random walk that started at location $\boldsymbol{x}$ at time $t = 0$. The operators $T_f$ and $T_b$ are thus the continuous analogues of the left and right multiplication by the finite matrix $M$.

We now take this analysis one step further and consider the limit $\varepsilon \to 0$. This is possible, since when $n = \infty$ each data point contains an infinite number of nearby neighbors. In this limit, since $\varepsilon$ also has the interpretation of a time step, the random walk converges to a diffusion process, whose probability density evolves continuously in time, according to

$$\frac{\partial p(x,t)}{\partial t} = \lim_{\varepsilon\to 0}\frac{p(x,t+\varepsilon)-p(x,t)}{\varepsilon} = \lim_{\varepsilon\to 0}\frac{T_f - I}{\varepsilon}p(x,t) \tag{15}$$

in which case it is customary to study the infinitesimal generators (propagators)

$$\mathcal{H}_f = \lim_{\varepsilon\to 0}\frac{T_f - I}{\varepsilon}, \qquad \mathcal{H}_b = \lim_{\varepsilon\to 0}\frac{T_b - I}{\varepsilon} \tag{16}$$

Clearly, the eigenfunctions of $T_f$ and $T_b$ converge to those of $\mathcal{H}_f$ and $\mathcal{H}_b$, respectively.

As shown in [11], the backward generator is given by the following Fokker-Planck operator

$$\mathcal{H}_b\psi = \Delta\psi - 2\nabla\psi\cdot\nabla U \tag{17}$$

which corresponds to a diffusion process in a potential field of $2U(\boldsymbol{x})$

$$\dot{\boldsymbol{x}}(t) = -\nabla(2U) + \sqrt{2D}\dot{\boldsymbol{w}}(t) \tag{18}$$

where $\boldsymbol{w}(t)$ is standard Brownian motion in $p$ dimensions and $D$ is the diffusion coefficient, equal to one in equation (17). The Langevin equation (18) is a common model to describe stochastic dynamical systems in physics, chemistry and biology [19, 20]. As such, its characteristics as well as those of the corresponding FP equation have been extensively studied, see [19]-[22] and many others. The term $\nabla\psi\cdot\nabla U$ in (17) is interpreted as a *drift* term towards low energy (high-density) regions, and as discussed in the next section, may play a crucial part in the definition of clusters.

Note that when data is uniformly sampled from $\Omega$, $\nabla U = 0$ so the drift term vanishes and we recover the Laplace-Beltrami operator on $\Omega$. The connection between the discrete matrix $M$ and the (weighted) Laplace-Beltrami or Fokker-Planck operator, as well as rigorous convergence proofs of the eigenvalues and eigenvectors of $M$ to those of the integral operator $T_b$ or infinitesimal generator $\mathcal{H}_b$ were considered in many recent works [4, 23, 17, 9, 24]. However, it seems that the important issue of boundary conditions was not considered.

Since (17) is defined in the bounded domain $\Omega$, the eigenvalues and eigenfunctions of $\mathcal{H}_b$ depend on the boundary conditions imposed on $\partial\Omega$. As shown in [9], in the limit $\varepsilon \to 0$, the random walk satisfies reflecting boundary conditions on $\partial\Omega$, which translate into

$$\frac{\partial\psi(\boldsymbol{x})}{\partial\boldsymbol{n}}\Big|_{\partial\Omega} = 0 \tag{19}$$

Table 1: Random Walks and Diffusion Processes

| Case | Operator | Stochastic Process |
|---|---|---|
| $\varepsilon > 0$ <br> $n < \infty$ | finite $n \times n$ <br> matrix $M$ | R.W. discrete in space <br> discrete in time |
| $\varepsilon > 0$ <br> $n \to \infty$ | operators <br> $T_f, T_b$ | R.W. in continuous space <br> discrete in time |
| $\varepsilon \to 0$ <br> $n = \infty$ | infinitesimal <br> generator $\mathcal{H}_f$ | diffusion process <br> continuous in time & space |

where $\boldsymbol{n}$ is a unit normal vector at the point $\boldsymbol{x} \in \partial\Omega$.

To conclude, the left and right eigenvectors of the finite matrix $M$ can be viewed as discrete approximations to those of the operators $T_f$ and $T_b$, which in turn can be viewed as approximations to those of $\mathcal{H}_f$ and $\mathcal{H}_b$. Therefore, if there are enough data points for accurate statistical sampling, the structure and characteristics of the eigenvalues and eigenfunctions of $\mathcal{H}_b$ are similar to the corresponding eigenvalues and discrete eigenvectors of $M$. For convenience, the three different stochastic processes are shown in table 1.

## 4 Fokker-Planck eigenfunctions and spectral clustering

According to (16), if $\lambda_\varepsilon$ is an eigenvalue of the matrix $M$ or of the integral operator $T_b$ based on a kernel with parameter $\varepsilon$, then the corresponding eigenvalue of $\mathcal{H}_b$ is $\mu \approx (\lambda_\varepsilon - 1)/\varepsilon$. Therefore the largest eigenvalues of $M$ correspond to the smallest eigenvalues of $\mathcal{H}_b$. These eigenvalues and their corresponding eigenfunctions have been extensively studied in the literature under various settings. In general, the eigenvalues and eigenfunctions depend both on the geometry of the domain $\Omega$ and on the profile of the potential $U(\boldsymbol{x})$. For clarity and due to lack of space we briefly analyze here two extreme cases. In the first case $\Omega = \mathbb{R}^p$ so geometry plays no role, while in the second $U(\boldsymbol{x}) = const$ so density plays no role. Yet we show that in both cases there can still be well defined clusters, with the unifying probabilistic concept being that the mean exit time from one cluster to another is much larger than the characteristic equilibration time inside each cluster.

**Case I:** Consider diffusion in a smooth potential $U(\boldsymbol{x})$ in $\Omega = \mathbb{R}^p$, where $U$ has a few local minima, and $U(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$ fast enough so that $\int e^{-U} d\boldsymbol{x} = 1 < \infty$. Each such local minimum thus defines a metastable state, with transitions between metastable states being relatively rare events, depending on the barrier heights separating them. As shown in [21, 22] (and in many other works) there is an intimate connection between the smallest eigenvalues of $\mathcal{H}_b$ and mean exit times out of these metastable states. Specifically, in the asymptotic limit of small noise $D \ll 1$, exit times are exponentially distributed and the first non-trivial eigenvalue (after $\mu_0 = 0$) is given by $\mu_1 = 1/\bar{\tau}$ where $\bar{\tau}$ is the mean exit time to overcome the highest potential barrier on the way to the deepest potential well. For the case of two potential wells, for example, the corresponding eigenfunction is roughly constant in each well with a sharp transition near the saddle point between the wells. In general, in the case of $k$ local minima there are asymptotically only $k$ eigenvalues very close to zero. Apart from $\mu_0 = 0$, each of the other $k - 1$ eigenvalues corresponds to the mean exit time from one of the wells into the deepest one, with the corresponding eigenfunctions being almost constant in each well. Therefore, for a finite dataset the presence of only $k$ eigenvalues close to 1 with a *spectral gap*, e.g. a large difference between $\lambda_k$ and $\lambda_{k+1}$ is indicative of $k$ well defined *global* clusters. In figure 1 (left) an example of this case is shown, where $p(\boldsymbol{x})$ is the sum of two well separated Gaussian clouds leading to a double well potential. Indeed there are only two eigenvalues close or equal to 1 with a distinct spectral gap and the first eigenfunction being almost piecewise constant in each well.

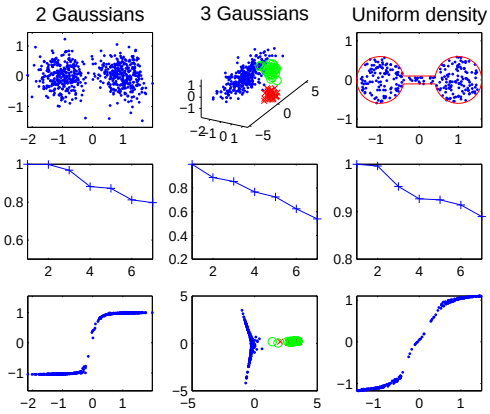


Figure 1: Diffusion map results on different datasets. Top - the datasets. Middle - the eigenvalues. Bottom - the first eigenvector vs. $\boldsymbol{x}_1$ or the first and second eigenvectors for the case of three Gaussians.

In stochastic dynamical systems a spectral gap corresponds to a separation of time scales between long transition times from one well or metastable state to another as compared to short equilibration times inside each well. Therefore, clustering and identification of metastable states are very similar tasks, and not surprisingly algorithms similar to the normalized graph Laplacian have been independently developed in the literature [25].

The above mentioned results are asymptotic in the small noise limit. In practical datasets, there can be clusters of different scales, where a global analysis with a single $\varepsilon$ is not suitable. As an example consider the second dataset in figure 1, with three clusters. While the first eigenvector distinguishes between the large cluster and the two smaller ones, the second eigenvector captures the equilibration inside the large cluster instead of further distinguishing the two small clusters. While a theoretical explanation is beyond the scope of this paper, a possible solution is to choose a location dependent $\varepsilon$, as proposed in [26].

**Case II:** Consider a uniform density in a region $\Omega \subset \mathbb{R}^3$ composed of two large containers connected by a narrow circular tube, as in the top right frame in figure 1. In this case $U(\boldsymbol{x}) = const$, so the second term in (17) vanishes. As shown in [27], the second eigenvalue of the FP operator is extremely small, of the order of $a/V$ where $a$ is the radius of the connecting tube and $V$ is the volume of the containers, thus showing an interesting connection to the Cheeger constant on graphs. The corresponding eigenfunction is almost piecewise constant in each container with a sharp transition in the connecting tube. Even though in this case the density is uniform, there still is a spectral gap with two well defined clusters (the two containers), defined entirely by the geometry of $\Omega$. An example of such a case and the results of the diffusion map are shown in figure 1 (right).

In summary the eigenfunctions and eigenvalues of the FP operator, and thus of the corresponding finite Markov matrix, depend on both geometry and density. The diffusion distance and its close relation to mean exit times between different clusters is the quantity that incorporates these two features. This provides novel insight into spectral clustering algorithms, as well as a theoretical justification for the algorithm in [13], which defines clusters according to mean travel times between points on the graph. A similar analysis could also be applied to semi-supervised learning based on spectral methods [28]. Finally, these eigenvectors may be used to design better search and data collection protocols [29].

**Acknowledgments:** The authors thank Mikhail Belkin and Partha Niyogi for interesting discussions. This work was partially supported by DARPA through AFOSR.

## References


[1] B. Schölkopf, A. Smola and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* 10, 1998.

[2] Y. Weiss. Segmentation using eigenvectors: a unifying view. *ICCV* 1999.

[3] J. Shi and J. Malik. Normalized cuts and image segmentation, *PAMI*, Vol. 22, 2000.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering, NIPS Vol. 14, 2002.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation* 15:1373-1396, 2003.

[6] A.Y. Ng, M. Jordan and Y. Weiss. On spectral clustering, analysis and an algorithm, NIPS Vol. 14, 2002.

[7] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, Proceedings of the $20^{th}$ international conference on machine learning, 2003.

[8] M. Saerens, F. Fouss, L. Yen and P. Dupont, The principal component analysis of a graph and its relationships to spectral clustering. ECML 2004.

[9] R.R. Coifman, S. Lafon, Diffusion Maps, to appear in Appl. Comp. Harm. Anal.

[10] R.R. Coifman & *al.*, Geometric diffusion as a tool for harmonic analysis and structure definition of data, parts I and II, *Proc. Nat. Acad. Sci.*, 102(21):7426-37 (2005).

[11] B. Nadler, S. Lafon, R.R. Coifman, I. G. Kevrekidis, Diffusion maps, spectral clustering, and the reaction coordinates of dynamical systems, to appear in Appl. Comp. Harm. Anal., available at http://arxiv.org/abs/math.NA/0503445.

[12] M. Meila, J. Shi. A random walks view of spectral segmentation, *AI and Statistics*, 2001.

[13] L. Yen L., Vanvyve D., Wouters F., Fouss F., Verleysen M. and Saerens M. , Clustering using a random-walk based distance measure. ESANN 2005, pp 317-324.

[14] N. Tishby, N. Slonim, Data Clustering by Markovian Relaxation and the information bottleneck method, NIPS, 2000.

[15] S. Yu and J. Shi. Multiclass spectral clustering. ICCV 2003.

[16] Y. Bengio et. al, Learning eigenfunctions links spectral embedding and kernel PCA, *Neural Computation*, 16:2197-2219 (2004).

[17] U. von Luxburg, O. Bousquet, M. Belkin, On the convergence of spectral clustering on random samples: the normalized case, NIPS, 2004.

[18] S. Lafon, A.B. Lee, Diffusion maps: A unified framework for dimension reduction, data partitioning and graph subsampling, submitted.

[19] C.W. Gardiner, *Handbook of stochastic methods*, third edition, Springer NY, 2004.

[20] H. Risken, *The Fokker Planck equation*, 2nd edition, Springer NY, 1999.

[21] B.J. Matkowsky and Z. Schuss, Eigenvalues of the Fokker-Planck operator and the approach to equilibrium for diffusions in potential fields, *SIAM J. App. Math.* 40(2):242-254 (1981).

[22] M. Eckhoff, Precise asymptotics of small eigenvalues of reversible diffusions in the metastable regime, *Annals of Prob.* 33:244-299, 2005.

[23] M. Belkin and P. Niyogi, Towards a theoeretical foundation for Laplacian-based manifold methods, COLT 2005 (to appear).

[24] M. Hein, J. Audibert, U. von Luxburg, From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians, COLT 2005 (to appear).

[25] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, F. Cordes, From simulation data to conformational ensembles, structure and dynamics based methods, *J. Comp. Chem.* 20:1760-74, 1999.

[26] L. Zelnik-Manor, P. Perona, Self-Tuning spectral clustering, NIPS, 2004.

[27] A. Singer, Z. Schuss, D. Holcman and R.S. Eisenberg, narrow escape, part I, submitted.

[28] D. Zhou & *al.*, Learning with local and global consistency, NIPS Vol. 16, 2004.

[29] I.G. Kevrekidis, C.W. Gear, G. Hummer, Equation-free: The computer-aided analysis of complex multiscale systems, *Aiche J.* 50:1346-1355, 2004.