Measure-based diffusion kernel methods

Amit Bermanis
School of Computer Science,
Tel Aviv University,
Tel Aviv 69978, Israel
Email: amitberm@post.tau.ac.il

Guy Wolf
School of Computer Science,
Tel Aviv University,
Tel Aviv 69978, Israel
Email: guy.wolf@cs.tau.ac.il

Amir Averbuch
School of Computer Science,
Tel Aviv University,
Tel Aviv 69978, Israel
Email: amir@math.tau.ac.il

***Abstract*—A commonly used approach for analyzing massive high dimensional datasets is to utilize diffusion-based kernel methods. The kernel in these methods is based on a Markovian diffusion process, whose transition probabilities are determined by local similarities between data points. When the data lies on a low dimensional manifold, the diffusion distances according to this kernel encompass the geometry of the manifold. In this paper, we present a generalized approach for defining diffusion-based kernels by incorporating measure-based information, which represents the density or distribution of the data, together with its local distances. The generalized construction does not require an underlying manifold to provide a meaningful kernel interpretation but assumes a more relaxed assumption that the measure and its support are related to a locally low dimensional nature of the analyzed phenomena.**

I. Introduction

The diffusion maps (DM) method [3] is a popular kernel method that utilizes a stochastic diffusion process to analyze the data. It defines diffusion affinities via symmetric conjugation of a transition probability operator. These probabilities are based on local distances between the data points. The Euclidean distances in the embedded space represent the diffusion distances in the original space. When the data is sampled from a low dimensional manifold, the diffusion paths follow the manifold and the diffusion distances capture its geometry.

In this paper, we enhance the DM method by incorporating information about the distribution of the data, in addition to local distances on which DM is based. This distribution is expressed in term of a measure over the observable space. The measure (and its support) replace the manifold assumption. We assume that the measure quantifies the likelihood for the presence of data over the geometry of the space. This assumption is significantly less restrictive than the need to have a manifold present. In practice this measure can either be provided as an input (e.g., by a-priori knowledge or a statistical model), or deduced from a given training set (e.g., by a density estimator). The manifold assumption can be expressed in terms of the measure assumption by setting the measure to be concentrated around an underlying manifold or (in the extremely restrictive case), to be supported by the manifold. Therefore, the measure assumption is not only less restrictive than the manifold assumption but it also generalizes it.

In the suggested construction, the used measure, which can represent densities, is separated from the distances and from the analyzed dataset. Therefore, when dealing with discrete data, this construction can utilize two different sets of samples: the analyzed dataset and the measure-related set with attached empirical measure values. Furthermore, from theoretical point of view, this construction combines continuous measures with either discrete or continuous datasets.

II. Problem setup

Let $\Omega \subseteq \mathbb{R}^n$, for some natural n, be a metric space with the Euclidean distance metric $\|\cdot\|$. The integration notation $\int \cdot dy$ in this paper will refer to the Lebesgue integral $\int_\Omega \cdot dy$ over the subspace Ω, instead of the whole space $\mathbb{R}^n$. Let μ be a probability measure defined on Ω and let $q(x)$ be the distribution function of μ, i.e., $d\mu(x) = q(x)dx$. This measure represents the distribution of data in Ω. We aim to combine the distance metric of Ω and the measure μ to define a kernel function $k(x,y)$, $x,y \in \Omega$, which represents the affinities between data points in Ω. Then, these affinities can be used to construct a diffusion map, as described in Section II-A, and utilize it to embed the data into a low-dimensional representation that considers both proximities and distributions of the data points.

A. Diffusion maps

The diffusion maps (DM) framework utilizes a set of affinities to define a Markovian (random-walk) diffusion process over the analyzed data [3]. The spectral properties of this process are then used to obtain a representation of the data, where diffusion distances are expressed as Euclidean distances. The achieved representation reveals the underlying patterns of the data such as clusters and differences between normal and abnormal regions.

Technically, DM is based on an affinity kernel k and the associated integral operator that is defined as $Kf(x) = \int k(x,y)f(y)dy$, $x \in \Omega$, for any function $f \in L^2(\Omega)$. The affinity kernel k is normalized by a set of degrees $\nu(x) \triangleq \int k(x,y)dy$, $x \in \Omega$, to obtain the transition probabilities $p(x,y) \triangleq k(x,y)/\nu(x)$, from $x \in \Omega$ to $y \in \Omega$, of the Markovian diffusion process. Under mild conditions on the kernel k, the resulting transition probability operator has a discrete decaying spectrum of eigenvalues $1 = \lambda_0 \geq |\lambda_1| \geq |\lambda_2| \geq \ldots$, which are used together with their corresponding eigenvectors $\vec{1} = \phi_0, \phi_1, \phi_2, \ldots$ to achieve the diffusion map of the data.

Each data point $x \in \Omega$ is embedded by this diffusion map to the diffusion coordinates $(\lambda_1\phi_1(x), \ldots, \lambda_\delta(x)\phi_\delta(x))$, where the exact value of δ depends on the spectrum of the transition probabilities operator P, whose kernel is $p(x, y)$. The relation between the diffusion distance metric $\|p(x, \cdot) - p(y, \cdot)\|$ and the Euclidean distances in the embedded space, is a result of the spectral theorem [3], [5]. When the data in Ω lies on a low dimensional manifold, its tangent spaces can be utilized to express the infinitesimal generator of the associated diffusion process in terms of the Laplacian operators on the manifold.

III. Measure-based diffusion and affinity kernels

In this section, we define and analyze an affinity kernel that is based on the distances in Ω and on the measure μ. We use this kernel together with the DM method, which was briefly described in Section II-A, to obtain a measure-based diffusion affinity kernel and its resulting diffusion map. In Section III-A, we show the relations between the infinitesimal generator of the resulting diffusion operator and the Laplacian operator on the space Ω and the measure μ.

In order to define the desired kernel, we first define the function

$$g_\varepsilon(t) \triangleq \begin{cases} e^{-t^2/\varepsilon} & t \le \rho\sqrt{\varepsilon} \\ 0 & \text{otherwise} \end{cases}, \qquad \text{(III.1)}$$

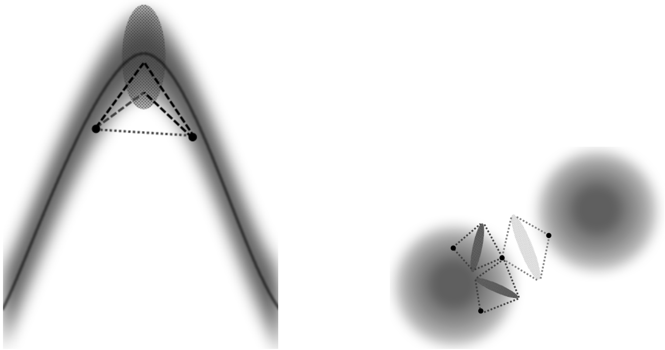
for any $\varepsilon > 0$ and some constant $\rho \gg 1$. Notice that for a sufficiently large ρ, the Gaussian kernel, which is usually used in the DM method, can be defined as $k_\varepsilon(x, y) \triangleq g_{2\varepsilon}(\|x - y\|)$, and this definition will be used in the rest of the paper. Definition III.1 uses the function g_ε to define an alternative kernel that incorporates both local distance information, as the Gaussian kernel does, and measure information, which the Gaussian kernel lacks.

Definition III.1 (Measure-based Gaussian Correlation kernel)**.** *The Measure-based Gaussian Correlation (MGC) affinity function* $\tilde{k}_\varepsilon : \Omega \times \Omega \to \mathbb{R}$ *is defined as* $\tilde{k}_\varepsilon(x, y) \triangleq \int g_\varepsilon(\|x - r\|) \cdot g_\varepsilon(\|y - r\|) d\mu(r)$. *The MGC integral operator is defined by this function as* $\tilde{K}_\varepsilon f(x) = \int \tilde{k}_\varepsilon(x, y) f(y) dy$ *for every function* $f \in L^2(\Omega)$ *and data point* $x \in \Omega$.

The MGC affinity from Definition III.1, is in fact the inner product in $L^2(\Omega, \mu)$ (correlation) between two Gaussians of width ε that are centered at x and y, respectively. This affinity takes into consideration the measure μ, between the described Gaussians around at the examined data points. The numerically significant positions of r in this correlation must be close enough to x and to y (based on their Gaussians of radius ε), but they must also be in an area with a high enough concentration of the measure μ. Notice that the measure information is considered and incorporated in the affinity definitions. From the identity $\|x - r\|^2 + \|y - r\|^2 = \frac{1}{2}\|x - y\|^2 + 2\left\|\frac{x+y}{2} - r\right\|^2$, the MGC affinity function becomes

$$\tilde{k}_\varepsilon(x, y) = k_\varepsilon(x, y) \cdot \int g_{\varepsilon/2}\left(\left\|\frac{x + y}{2} - r\right\|\right) d\mu(r). \qquad \text{(III.2)}$$

Equation III.2 shows the relation between the MGC kernel and the Gaussian kernel $k_\varepsilon(x, y)$. While the Gaussian affinity only considers the distances between the examined data points, the MGC affinity also considers the region in which this distance is measured by using a Gaussian around the midpoint between them. This midpoint represents the direct path that determines the distance between the two data points. For a given distance between two data points, the MGC affinity increases when its path lies in an area with a high concentration of the measure μ, and decreases when it lies in an area with a low concentration of μ. If the measure μ is uniform over Ω, then the MGC kernel becomes the same as the Gaussian kernel up to a constant.



(a) When the data lies around a curve, the MGC affinities consider paths that follow the curve.

(b) When the data lies in two separate clusters, the affinities between data points within a cluster are higher than data points from a different cluster.

Fig. III.1. An illustration of the MGC affinities in two common data analysis scenarios. For every pair of compared data points, the significant values of the integration variable r, from Definition III.1 or the equivalent representation from Eq. III.2, are marked.

The dual representation of the MGC kernel in Definition III.1 and Eq. III.2 can be used to detect and consider several common patterns in data analysis directly from the initial construction of the kernel. Figure III.1(a) uses the formulation in Definition III.1 to illustrate a case when the data is concentrated in areas around a curve with significant curvatures. In this case, the affinity will be more affected by the distances over the path that follows the “noisy” curve and not by the directions that follow sparse areas and bypass the curve. Figure III.1(b) uses the formulation in Eq. III.2 to illustrate the affinities when the data is concentrated in two distinct clusters. In this case, we can see that the affinity between data points from different clusters is significantly reduced due to the measure even if they are relatively close.

As proved in [1], the presented MGC affinity kernel satisfies the spectral properties that are required (and assumed) in [3], [5] for its utilization with the DM framework. These properties enable us to define a diffusion process that is based on the MGC affinities. Then, the resulting diffusion map is used to embed the data in a way that considers the distances and the measure distribution.

A. *Infinitesimal generator*

The DM framework is based on Markovian diffusion process, which is defined and represented by a transition probability operator denoted by P_ε. The infinitesimal generator of this operator encompasses the nature of the diffusion process. In [3], [5], it was shown that when the data is sampled from a low dimensional underlying manifold, the infinitesimal generator of P_ε has the form of *Laplacian+Potential*. In this section, we show a similar result, when using the MGC-based diffusion without requiring the underlying manifold assumption to hold.

The MGC affinity function $\tilde{k}_\varepsilon$ is symmetric and positive, i.e., $\tilde{k}_\varepsilon(x, y) > 0$ for any pair of data points $x, y \in \Omega$. To convert it to be a transition kernel of a Markov chain on Ω, we normalize it to be $\tilde{p}_\varepsilon(x, y) \triangleq \frac{\tilde{k}_\varepsilon(x,y)}{\nu_\varepsilon(x)}$. We define the corresponding stochastic operator $\tilde{P}_\varepsilon f(x) \triangleq \int \tilde{p}_\varepsilon(x, y) f(y) dy$.

The infinitesimal generator of the diffusion transition operator $\tilde{P}_\varepsilon$ is defined as $\mathcal{L} \triangleq \lim_{\varepsilon \to 0}(I - \tilde{P}_\varepsilon)/\varepsilon$. Theorem III.1, whose proof appears in [1], shows that the operator $\mathcal{L}$ takes the form *Laplacian+potential*, which is similar to the result shown in [5, Corollary 2]. The expression, which Theorem III.1 provides for $\mathcal{L}$, characterizes the differential equation for diffusion processes [2], [4].

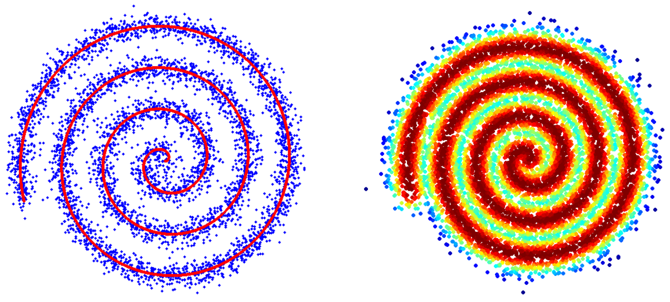
Theorem III.1. *If the density function q is in $C^4(\Omega)$, then the infinitesimal generator $\mathcal{L}$ of the MGC-based diffusion operator is*

$$\mathcal{L}f = -\frac{m_2}{m_0}\left(\Delta f + \left\langle \frac{\nabla q}{q}, \nabla f \right\rangle\right), \qquad f \in C^4(\Omega),$$

where, $m_0 = \int g_1(\|x\|)dx$ and $m_2 = \int g_1(\|x\|)(x^{(j)})^2 dx$.

IV. Geometric example

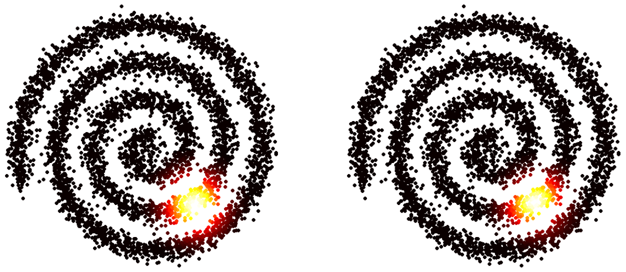
In this section, we demonstrate the MGC kernel and the resulting diffusion map. A noisy data that is spread around a spiral curve is analyzed, and the results are compared with the "classic" DM [3]. This example also demonstrates the separation between the analyzed data and the data distribution, which is a unique feature of the presented method.



(a) Noisy data around the curve

(b) An exponentially-decaying measure around the curve

Fig. IV.1. A spiral curve with 5000 noisy data points concentrated around it, and 10^4 points that represent an exponentially-decaying measure around the curve. Red color indicates large measure weights and blue color indicates small measure weights.



(a) K neighborhood

(b) $\tilde{K}_v$ neighborhood

Fig. IV.2. A neighborhoods from the Gaussian kernel and the MGC kernels on the spiral curve. Close points are colored by white, and far points are colored by black.

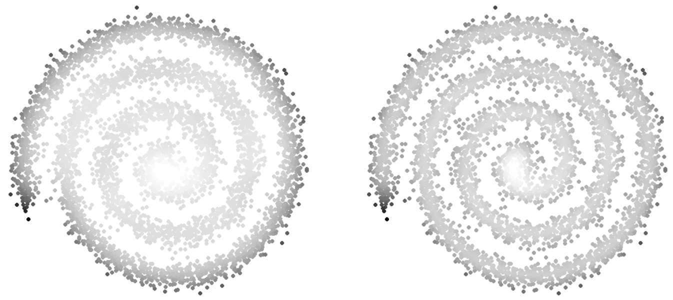
We use a noisy spiral curve (see Fig. IV.1(a)) for the comparison between MGC-based DM and the classical DM. The dataset was produced by sampling 500 equally spaced points from the curve and then sampling 10 normally distributed data points around each of these curve points. The resulting data has 5000 data points that lie in areas around the curve, as shown in Fig. IV.1(a), where the curve is marked in red and the noisy data points are marked in blue. We used the same scale meta-parameter ε to the compared DM applications. This meta-parameter was set to be sufficiently high to overcome the noise and to detect the high affinity between data points that originated from the same position (out of the 500 curve points) on the curve.

The MGC kernel from Definition III.1 requires to define a measure over the area where the data lies. Notice that the measure of the actual data points is not required. We can define a completely different set of points r from Definition III.1 and then define their weights, which represent their measure values. The measure we used is based on 10^4 points, distributed normally around a spiral curve. The weights of the point decay exponentially in relation to their distance from the curve. The resulting measure is denoted by μ_v and it is presented in Fig. IV.1(b).

We use the notation $\tilde{K}_v$ to denote the matrix that results from Definition III.1, with the measure μ_v. Notice that even though the measure is based on 10^4 positions of the integration variable r (from Definition III.1), the kernel and its normalized versions are of size 5000×5000, since the data has only 5000 data points.

Figure IV.2 compares the neighborhoods that are represented by the kernels K and $\tilde{K}_v$. While the Gaussian kernel captures inter-level affinities (i.e., it links different levels of the spiral), the MGC kernel only capture relations in the same level of the spiral, thus, it is able to separate between these levels. In addition, the shape of the neighborhoods of the MGC kernel form ellipses whose major axes clearly follow the significant tangential directions of the curve. The Gaussian kernel, however, captures circular neighborhoods that do not express any information about the significant directions of the data.

The embedding, which is achieved by DM, is based on



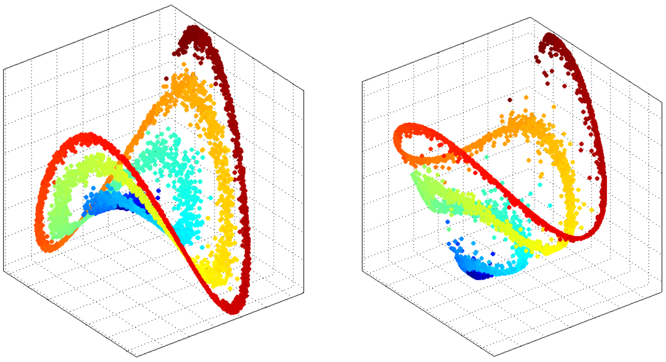
(a) Gaussian-based stationary distribution

(b) MGC-based stationary distribution

Fig. IV.3. The stationary distributions of: (a) the Gaussian-based diffusion process, and (b) the MGC-based diffusion process (low densities are represented by dark gray levels, and vise-versa.)

a diffusion process that has a stationary distribution when the time is taken to infinity. This distribution reveals the concentrations and the underlying potential of the diffusion process. It is represented by the first left eigenvector of the diffusion transition operator. Figure IV.3 compares the stationary distributions of the Gaussian-based diffusion with the MGC-based diffusion. This comparison shows that the Gaussian-based diffusion considers the entire spiral as one pit of potential. At infinity, the diffusion is distributed over the entire region of the curve. The MGC-based diffusion, on the other hand, separates different levels of the spiral. At infinity, this diffusion is concentrated on the curve levels themselves and not on the areas between them.

Finally, we compare between the embedded spaces of the Gaussian-based DM and the MGC-based DM. Figure IV.4 presents these spaces based on the first three diffusion coordinates. The comparison in Fig. IV.4 clearly shows that the MGC-based embedding results in a better separation between the spiral levels. Figure IV.4 further establishes this observation by showing that, in fact, the Gaussian-based diffusion considers the whole noisy spiral as a two-dimensional disk. The MGC-based embedding, on the other hand, separates the levels of the spiral by “stretching” it apart in the three-dimensional embedded space.



(a) Gaussian-based DM

(b) MGC-based DM

Fig. IV.4. The first three diffusion coordinates of the Gaussian-based and MGC-based DM embeddings.

The superior results (e.g., separation between the spiral levels) of the MGC-based DM demonstrate its robustness to noise. The reason for this robustness is because the noise is part of the model on which the MGC construction is based. The Gaussian-based DM assumes that the data lies on (or it is sampled from) an underlying manifold, and any significant noise outside this manifold may violate this assumption. The MGC-based DM, on the other hand, already assumes variable concentrations and distributions of the data, which are represented by the measure and incorporated into the affinities. Therefore, this setting is more natural when dealing with data that is concentrated *around* an underlying manifold structure but does not necessarily lie on the manifold.

V. Conclusion

We presented a generalized version of DM, which is based on the MGC kernel instead of the Gaussian kernel. We replaced the commonly-used manifold assumption in DM with a measure assumption. Namely, we assume access to a measure that represents the locally low dimensional nature of the analyzed data, its distributions and its densities. The MGC kernel was presented and formulated in two equivalent forms that incorporate the measure-based information together with local distances between data points. The infinitesimal generator of the MGC-based diffusion process is similar to the diffusion process in [3], and its spectral properties enable its utilization for dimensionality reduction.

We demonstrated the robustness of the MGC-based DM to noise, which is due to the noise being considered as part of the measure assumption while it violates the manifold assumption. Since the MGC-based construction considers the measure and the data points separately, it is able to analyze a given measure distribution by using a separated grid, as we will show in future work. This application cannot be achieved by the classic DM [3], which is based solely on local distances and does not consider a separately-provided measure.

Acknowledgments

This research was supported by the Israel Science Foundation (Grant No. 1041/10) and the Eshkol Fellowship from the Israeli Ministry of Science & Technology.

References

[1] A. Bermanis, G. Wolf, and A. Averbuch. Diffusion-based kernel methods on Euclidean metric measure spaces. *Submitted*, 2012.

[2] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling and Simulation*, pages 842–864, 2008.

[3] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[4] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

[5] S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, May 2004.