Random Diffusion Representations

Moshe Salhov
School of Computer Science
Tel Aviv University
Tel Aviv, Israel
Email: moshebar-s@013.net

Amir Averbuch
School of Computer Science
Tel Aviv University
Tel Aviv, Israel
Email: amir1@tauex.tau.ac.il

***Abstract*—The diffusion maps is a kernel based method for manifold learning and data analysis that models a Markovian process over data. Analysis of this process provides meaningful information about the inner geometric structures in the data.**

In this paper, we present a representation framework for analyzing datasets. This framework is based on a random approximation of the diffusion maps kernel. The resulted representation approximate the pair-wise diffusion distance, does not depend on the data size and it is invariant to scale.

I. Introduction

Kernel methods constitute a wide class of algorithms for nonparametric data analysis of massive high dimensional datasets. Typically, a limited set of underlying factors generates the high dimensional observable parameters via nonlinear mappings. The nonparametric nature of these methods enables to uncover hidden structures in the data. These methods extend the well known MDS [4]method. They are based on an affinity kernel construction that encapsulates the relations (distances, similarities or correlations) among multidimensional data points. Spectral analysis of this kernel provides an efficient representation of the data that simplifies its analysis. Methods such as Isomap [15], LLE [9], Laplacian eigenmaps [1], Hessian eigenmaps [6] and local tangent space alignment,extend the MDS paradigm by considering the manifold assumption. Under this assumption, the data is assumed to be sampled from a low intrinsic dimensional manifold that captures the dependencies between the observable parameters. The corresponding spectral embedding spaces in these methods preserve the geometry of the manifold, which incorporates the underlying factors of the data.

The diffusion maps (DM) method [3] is a kernel-based method that models and analyzes a Markovian process over the data. It defines a transition probability operator based on local affinities between the multidimensional data points. By spectral decomposition of this operator, the data is embedded into a low dimensional Euclidean space, where distances 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.

The DM embedding was utilized in a wide variety data and pattern analysis techniques. For example it was used to improve audio quality by suppressing transient interference [14]. It was utilized in [12] for detecting moving vehicles. Additionally, gene expression analysis [10] and source localization [13]. Furthermore, the DM method can be utilized for fusing different sources of data [7].

In general, kernel methods can provide a representation of the given data via a spectral decomposition. However, this representation changes as the data size grows and prone to reordering and sign change inflicted from the spectral decomposition. Furthermore, the required computational complexity, which is dictated by the spectral decomposition, is $O(n^3)$ that is not feasible for a very large dataset. Recently, a closed form representation was developed for the measure-based Gaussian correlation (MGC) kernel in [2], [11].

In this paper, we extend the result from [11] to compute a representation that preserves the diffusion distances between data points based on the DM framework [3].This representation is applicable for very large datasets. It utilizes a Markovian diffusion process to define and represent nonlinear relations between data points. It provides a diffusion distance metric that correlates with the intrinsic geometry of the data. The suggested representation is invariant to the dataset size and cost $O(1)$ operations per a given data point.

II. Problem formulation and mathematical preliminaries

Consider a big dataset $X \subseteq \mathbb{R}^m$ such that for any practical purposes the size of X is considered to be infinite. Without-loss-of-generality, we assume that for all $x \in X$, $\|x\| \leq 1$. Implementation of a kernel method, which uses a full spectral decomposition, becomes impractical when the dataset size is big. Instead, we suggest to represent the given dataset via the density of data points in it using the standard DM kernel. In other words, let $q : \mathbb{R}^m \to [0, 1]$ be the density function of X. We aim to find an explicit embedding function denoted by $f_q : \mathbb{R}^m \to \mathbb{R}^k$, where k is the embedding dimension such that $m \ll k$. Each member in f_q, which is denoted by $f_q(x)$, $x \in X$, depends on the density function q.

DM provides a multiscale view of the data via a family of geometries that are referred to by *diffusion geometries*. Each geometry is defined by both the associated diffusion metric and the diffusion time parameter t that are linked by $d_\varepsilon^{(t)} : X \times X \to \mathbb{R}_+$ where ε is a localization parameter. The *diffusion maps* are the associated functions $\Psi^{(t)} : X \to \mathbb{R}^k$ that embed the data into Euclidean spaces, where the diffusion geometries are preserved such that $\|\Psi^{(t)}(x) - \Psi^{(t)}(y)\| \approx d_\varepsilon^{(t)}(x,y)$, $x, y \in X$.

Given an accuracy requirement $\zeta > 0$, we aim to design an embedding f_q that preserves the diffusion geometry for $t = 1$ such that for all $x, y \in X$,

$$\left| \|f_q(x) - f_q(y)\| - d_\varepsilon^{(1)}(x,y) \right| \leq \zeta. \tag{II.1}$$

We call the embedding f_q the diffusion representation. From the requirement in Eq. (II.1), the Euclidian distance between pairs of representatives approximates the diffusion distance between the corresponding data points over the density of these data points in the DM kernel when $t = 1$. If Eq. (II.1) holds, then f_q preserves the diffusion geometry of the dataset in this sense.

The rest of this section is dedicated to provide additional details regarding the diffusion geometries that utilize the DM kernel.

A. Diffusion geometries

A family of diffusion geometries of a measurable space (X, μ) with a measure μ is determined by imposing a Markov process over the space. Given a non-negative symmetric kernel function $k_\varepsilon : X \times X \to \mathbb{R}_+$, then an associated Markov process over the data via the stochastic kernel $p_\varepsilon : X \times X \to \mathbb{R}_+$ is

$$p_\varepsilon(x,y) \triangleq k_\varepsilon(x,y)/\nu_\varepsilon(x), \tag{II.2}$$

where $\nu_\varepsilon : X \to \mathbb{R}$ is the local volume function. In a discrete setting, it is called the degree function. In a continues settings, the local volume function is defined by

$$\nu_\varepsilon(x) \triangleq \int_X k_\varepsilon(x,y) d\mu(y). \tag{II.3}$$

The associated Markovian process over X is defined via the conjugate operator of the integral operator $Pq(x) = \int_X p_\varepsilon(x,y)q(y)d\mu(y)$ that is denoted by P^*. Thus, for any initial probability distribution q_0 over X, $q_1 = P^* q_0$ is the probability distribution over X after a single time step. The probability distribution over X after t time steps is given by the t-th power of P^*. Specifically, if the initial probability measure is concentrated in a specific data point $x \in X$, i.e. $q_0 = \delta(x)$, then the probability distribution after t time steps is $(P^*)^t \delta(x)$, denoted also by $p_\varepsilon^{(t)}(x, \cdot)$. Thus, $p_\varepsilon^{(t)}(x,y)$ is the probability that a random walker, which started his walk in $x \in X$, will end in $y \in X$ after t time steps. Based on this, the t-time diffusion geometry is defined by the distances between probability distributions such that for all $x, y \in X$

$$d_\varepsilon^{(t)}(x,y) \triangleq \left\| p_\varepsilon^{(t)}(x,\cdot) - p_\varepsilon^{(t)}(y,\cdot) \right\|_{L^2(\mathbb{R}^m)}. \tag{II.4}$$

Equations II.1 and II.4 suggest that the embedding $f_q(x)$ approximately preserves (for $t = 1$) the distance between probability distributions. Such a family of geometries can be defined for any Markovian process and not necessarily for a diffusion process. It is proved in [3] that under specific conditions, the defined Markovian process approximates the diffusion over a manifold from which the dataset X is sampled. If the Markovian process is ergodic, then it has a unique probability distribution $\hat{\nu}_\varepsilon : X \to \mathbb{R}_+$, to which it converges independently of its initial distribution, namely, for any $y \in X$, $\hat{\nu}_\varepsilon(y) = \lim_{t\to\infty} p_\varepsilon^{(t)}(x,y)$, independently of x. This probability measure is an L_1 normalization of the local volume function (Eq. (II.3)), i.e. $\hat{\nu}_\varepsilon(y) = \nu_\varepsilon(y) / \int_X \nu_\varepsilon(y) d\mu(y)$.

B. Measure-based DM kernel

Mathematically, for the analyzed domain $X \subset \mathbb{R}^m$ and for the measure domain $M \subset \mathbb{R}^m$ with a density function $q : M \to \mathbb{R}_+$ defined on the measure domain, the DM kernel $k_\varepsilon : X \times X \to \mathbb{R}_+$ is defined as

$$k_\varepsilon(x,y) \triangleq g_m(r; x, \varepsilon I_m), \tag{II.5}$$

where I_m is an $m \times m$ unit matrix. For a fixed mean vector $\theta \in \mathbb{R}^m$ and a covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$, $g_m(r;\theta,\Sigma) : \mathbb{R}^m \to \mathbb{R}_+$ is the normalized Gaussian function given by

$$g_m(r;\theta,\Sigma) \triangleq \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(r-\theta)^T \Sigma^{-1} (r-\theta) \right\}. \tag{II.6}$$

Since the DM kernel in Eq. (II.5) is symmetric and positive, it can be utilized to establish a Markov process as was described in Section II-A. The associated diffusion parameters from Eqs. (II.2) , (II.3) and (II.4) are p_ε, ν_ε and $d_\varepsilon^{(t)}$, respectively.

III. Explicit forms for the diffusion distance and stationary distribution

In general, the integral in Eq. (II.5) does not have an explicit form. However, for our purposes, we adopt the Gaussian Mixture Model (GMM), which assumes that the density q is a superposition of normal distributions. Under this assumption, q takes the form

$$q(r) = \sum_{j=1}^n a_j g_m(r;\theta_j,\Sigma_j), \ \sum_{j=1}^n a_j = 1, \tag{III.1}$$

for appropriate mean vectors θ_j and covariance matrices Σ_j, $j = 1, \ldots, n$, (see Eq. (II.6)). Estimating Eq. (III.1) is a generally known problem that has been extensively investigated such as in [5], [8] with many published implementations. Such an estimation enables to provide an explicit (closed form) representation of the diffusion geometry in Eq. (II.4).

First, a closed form for the inner product $W_{x,z} = \langle p_\varepsilon^{(1)}(x,\cdot), p_\varepsilon^{(1)}(z,\cdot)\rangle_{L^2(\mathbb{R}^k)}$, $x, z \in X$, is presented. This inner product closed form enables to get an explicit formulation for the first time step of the DM distance $d_\varepsilon^{(1)}(x,z)$. This formalism is established in Theorem III.1.

Theorem III.1. *Assume that the GMM assumption in Eq.* (III.1) *holds. Then, for any $x \in X$, the stationary distribution $\nu_\varepsilon(x)$ have explicit forms given by*

$$\nu_\varepsilon(x) \ = \ \sum_{j=1}^n a_j g_m(x;\theta_j,\varepsilon I_m + \Sigma_j) \tag{III.2}$$

Proof. by definition of the stationary distribution we have

$$\begin{aligned} \nu_\varepsilon(x) &\triangleq \int_X k_\varepsilon(x,y)d\mu(y). \\ &= \int_X g_m(x;y,\varepsilon I_m)q(y)dy \\ &= \int_X g_m(x;y,\varepsilon I_m)\sum_{j=1}^n a_j g_m(y;\theta_j,\Sigma_j)dy \\ &= \sum_{j=1}^n \int_X g_m(x;y,\varepsilon I_m)a_j g_m(y;\theta_j,\Sigma_j)dy \\ &= \sum_{j=1}^n a_j g_m(x;\theta_j,\varepsilon I_m+\Sigma_j) \end{aligned} \tag{III.3}$$

□

Combination of Theorem III.1 with Eqs. (II.2), (II.3) and (II.4) formulate the first time step ($t=1$) diffusion metric as

$$d_\varepsilon^{(1)}(x,z) = \frac{W_{x,x}}{\nu_\varepsilon(x)\nu_\varepsilon(x)} + \frac{W_{z,z}}{\nu_\varepsilon(z)\nu_\varepsilon(z)} - \frac{2W_{x,z}}{\nu_\varepsilon(x)\nu_\varepsilon(z)}. \tag{III.4}$$

IV. Randomized Diffusion Maps of the Analyzed Domain

The diffusion distance provides a relation between pair of data points in the analyzed domain. In this section, we find a representation of any data point in the analyzed domain that preserves the diffusion distance relation.

let $W_{x,z}$ be the inner product $W_{x,y} \triangleq \langle k_\varepsilon(x,\cdot), k_\varepsilon(z,\cdot)\rangle_{L^2(\mathbb{R}^k)}$. Then, for any $x,z \in X$ we have,

$$\begin{aligned} W_{x,z} &\triangleq \int_X p_\varepsilon(x,y)p_\varepsilon(z,y)q(y)dy \\ &= \tfrac{1}{\nu_\varepsilon(x)\nu_\varepsilon(z)}\int_X k_\varepsilon(x,y)k_\varepsilon(z,y)q(y)dy. \end{aligned} \tag{IV.1}$$

However, the integral form in Eq. IV.1 can be reformulated as the expectation operator,

$$\begin{aligned} &\tfrac{1}{\nu_\varepsilon(x)\nu_\varepsilon(z)}\int_X k_\varepsilon(x,y)k_\varepsilon(z,y)q(y)dy \\ &\quad = \mathbb{E}_{q(y)}k_\varepsilon(x,y)k_\varepsilon(z,y), \end{aligned} \tag{IV.2}$$

where the expectation is over the stationary distribution q. Using the expectation in Eq. IV.2, the inner product $W_{x,z}$ can be approximated using a random sample from q as,

$$\begin{aligned} &\mathbb{E}_{q(y)}k_\varepsilon(x,y)k_\varepsilon(z,y) \\ &\approx \tfrac{1}{L}\sum_{l=1}^L k_\varepsilon(x,y_l)k_\varepsilon(z,y_l), \end{aligned} \tag{IV.3}$$

where y_i, $1 \le i \le L$ are L random samples (with distribution q) from X. The approximation in Eq. IV.3 can be reformulated as the inner product,

$$\approx \frac{1}{L}\sum_{l=1}^L k_\varepsilon(x,y_l)k_\varepsilon(z,y_l) = \phi(x)^T\phi(z), \tag{IV.4}$$

where $\phi(x) = \frac{1}{\sqrt{L}}[k_\varepsilon(x,y_1),...,k_\varepsilon(x,y_L)]$, $\quad y_l \sim q, \quad 1 \le l \le L$. The vector $f_q(x) = \phi(x)$ is the embedding of the data point x into a space that approximatly preserve the inner product $W_{x,z}$ and hence approximatly preserves the diffusion geometry.

The approximation error is given by,

$$\zeta \triangleq \left| \frac{1}{L}\sum_i^n p(x,y_i)p(z,y_i) - \int_X p(x,y)p(z,y)dy \right| \tag{IV.5}$$

. We would like to bound the probability of ζ to be larger than some confidence η. Introducing the sampled integral and its average into the Chebyshev's inequality gives

$$\begin{aligned} &\mathbb{P}(\zeta \ge \eta) \\ &\le \tfrac{\sigma(\frac{1}{L}\sum_i^L p(x,y_i)p(z,y_i))^2}{\eta^2}, \end{aligned} \tag{IV.6}$$

Using Bienaym formula for variance of the sum of uncorrelated samples we get

$$\mathbb{P}(\zeta \ge \eta) \le \frac{\sigma(p(x,y_i)p(z,y_i))^2}{L\eta^2}. \tag{IV.7}$$

Lets look at the term $p(x,y_i)p(z,y_i)$. This term by definition can formulated as

$$p(x,y_i)p(z,y_i) = \frac{1}{\nu_\varepsilon(x)}k_\varepsilon(x,y)\frac{1}{\nu_\varepsilon(z)}k_\varepsilon(z,y) \le 1. \tag{IV.8}$$

Hence, the variance is bounded by 1 and the first coarse bound we have is given by $\mathbb{P}(\ge k) \le \frac{1}{L\eta^2}$.

V. Experimental Results

This section demonstrate the principles of the DM closed-form embedding. The following example presents an analysis of a density function for which the stationary distribution is analytically known. The closed-form stationary distribution in this case is compared to the analytical stationary distribution.

Let the density function $q(r) \in \mathbb{R}^2$ includes two flat squares with probability $\frac{1}{5}$ to draw samples from the lower square and $\frac{4}{5}$ to draw samples from the upper square. In other words,

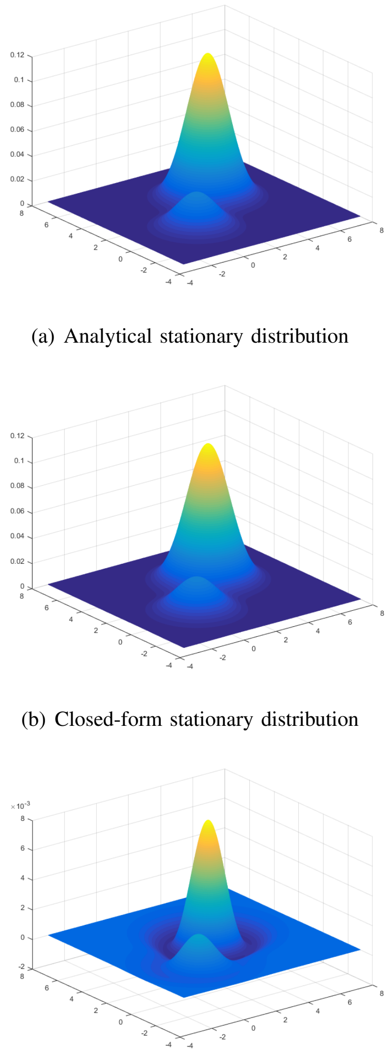
$$q\left(r\right) = \frac{1}{5}\chi_{[0,1]\times[0,1]}(r) + \frac{4}{5}\chi_{[3,4]\times[3,4]}(r) \tag{V.1}$$

where $\chi_{[a,b]\times[c,d]}$ is the indicator function for the square $abcd$. Eq. II.3 formulate the stationary distribution computation. Given $\varepsilon = 1$, the integration in Eq. II.3 can be analytically solved as

$$\nu\left(x_1,x_2\right) = 0.2H(0,1,x_1,x_2) + 0.8H(3,4,x_1,x_2), \tag{V.2}$$

where $H(a,b,x_1,x_2)$ is a given by $H(a,b,x_1,x_2) = \frac{1}{4}\left(\operatorname{erf}(b-x_1) - \operatorname{erf}(a-x_1)\right)\left(\operatorname{erf}(b-x_2) - \operatorname{erf}(a-x_2)\right)$, and $\operatorname{erf}(x)$ is the Gauss error function.

Given a properly trained GMM, the stationary distribution is computed using Eq. III.2. First, 2000 data points were randomized from the distribution in Eq. V.1. Then, A 1000×1000 grid was constructed to compute the stationary distribution via Eq. III.2 and based on the analytical solution in Eq. V.2.



(a) Analytical stationary distribution

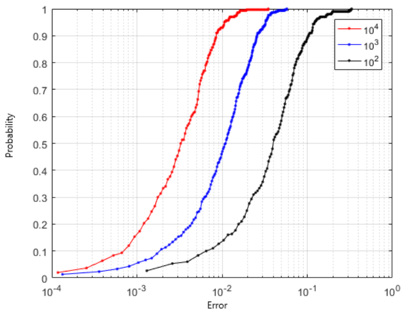
(b) Closed-form stationary distribution

(c) The error

Fig. V.1. Comparing between the closed-form stationary distribution and the analytical stationary distribution. (a) Analytical stationary distribution. (b) Closed-form stationary distribution. (c) The error.

Figure V.1(b) present the stationary distribution with minor distortion compared to Figure V.2(a). The error, which is presented in Figure V.1(c), is a result of the GMM training over a small set of data points. The difference is in the order of 5% and is the result of the GMM training error. The full paper will contain two more examples that demonstrate both the distance preservation and an application on real data.

Figure V.2 present the comulative distribution function (CDF) of the diffusion distance error between Eq. $II.4$ and its corresponding approximation that is based on introducing Eq. IV.4 into Eq. III.4. For the comparison we randomized $L = 10^2, 10^3, 10^4$ samples from X to find $\phi(x)$. The results in Fig. V.2 suggest that for each two order of magnitude increase in L an order of magnitude in accuracy is achieved. Hence, the proposed method is more appropriate where the diffusion distance is desired but low dimensionally is not necessary for the analysis.



(a) Diffusion distance Approximation error CDF

Fig. V.2. Diffusion distance approximation error CDF for $L = 10^2, 10^3, 10^4$

Acknowledgment

This research was partially supported by the Israel Science Foundation (ISF, 1556/17), Blavatnik Computer Science Research Fund and INdo-ISrael CollAborative for Infrastructure Security - Israel Ministry of Science and Technology 3-14481.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[2] A. Bermanis, G. Wolf, and A. Averbuch. Diffusion-based kernel methods on Euclidean metric measure spaces. *Applied and Computational Harmonic Analysis*, 2015. DOI:10.1016/j.acha.2015.07.005.

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

[4] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, London, UK, 1994.

[5] N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):pp. 463–474, 1969.

[6] D.L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100:5591–5596, May 2003.

[7] S. Lafon, Y. Keller, and R.R Coifman. Data fusion and multicue data matching by diffusion maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1784–1797, 2006.

[8] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[9] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, December 2000.

[10] X. Rui, S. Damelin, and D.C. Wunsch. Applications of diffusion maps in gene expression data-based cancer diagnosis analysis. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 4613–4616, 2007.

[11] Moshe Salhov, Amit Bermanis, Guy Wolf, and Amir Averbuch. Diffusion representations. *Applied and Computational Harmonic Analysis*, 45(2):324 – 340, 2018.

[12] A. Schclar, A. Averbuch, N. Rabin, V. Zheludev, and K. Hochman. A diffusion framework for detection of moving vehicles. *Digital Signal Processing*, 20(1):111 – 122, 2010.

[13] R. Talmon, I. Cohen, and S. Gannot. Supervised source localization using diffusion kernels. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 245–248, 2011.

[14] R. Talmon, I. Cohen, and S. Gannot. Single-channel transient interference suppression with diffusion maps. *IEEE transactions on audio, speech, and language processing*, 21(1-2):132–144, 2013.

[15] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.