Measure-based diffusion grid construction and high-dimensional data discretization

Amit Bermanis, Moshe Salhov, Guy Wolf, Amir Averbuch*

School of Computer Sciene, Tel Aviv University, Tel Aviv 69978, Israel

Abstract

The diffusion maps framework 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 concerning inner geometric structures in the data. Recently, it was suggested to replace the standard kernel by a measure based kernel, which incorporates information about the density of the data. Thus, the manifold assumption is replaced by a more general measure assumption.

The measure-based diffusion kernel utilizes two separate independent datasets. The first is the set by which the measure is determined. This measure correlates with a density that represents normal behaviors and patterns in the data. The second set consists of the analyzed data points that are embedded by the metastable states of the underlying diffusion process. This set can either be contiguous or discrete.

In this paper, we present a data discretization methodology for analyzing a contiguous domain. The obtained discretization is achieved by constructing a uniform grid over this domain. This discretization is designed to approximate the continuous measure-based diffusion process by a discrete random walk process. This paper provides a proved criterion to determine the grid resolution that ensures a controllable approximation error for the continuous steady states by the discrete ones. Finally, the presented methodology is demonstrated on analytically generated data.

Keywords: Dimensionality reduction, kernel PCA, diffusion-based kernel, measure-based information, grid construction, data discretization

1. 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

*Corresponding author, Tel: +972-54-5694455, Fax: +972-3-6422020
Email address: amir@math.tau.ac.il (Amir Averbuch)

mappings. The nonparametric nature of these methods enables to uncover hidden structures in the data. These methods extend the well known MDS [5, 14] 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 [21], LLE [19], Laplacian eigenmaps [1], Hessian eigenmaps [8] and local tangent space alignment [26, 27], 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 [4] is a kernel 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 measure-based Gaussian correlation (MGC) framework [2, 3] enhances the DM method by incorporating information about the distribution of the data in addition to the local distances on which DM is based. This distribution is modeled by a probability measure, which is assumed to quantify the likelihood of data presence over the geometry of the space. The measure and its support in this method generalize and replace the manifold assumption. Thus, the diffusion process is accelerated in high density areas of the data, rather than depending solely on a manifold geometry. As shown in [2], the compactness of the associated integral operator enables dimensionality reduction by utilization of the DM framework.

This measure-based construction consists of two independent sets of data points. The first set is the domain on which the measure is defined or, equivalently, the support of the measure. The second set is the domain on which the MGC kernel function and the resulting diffusion process are defined. These *measure domain* and *analyzed domain* may, in some cases, be identical, but separate sets can also be considered by the construction. The latter case enables the utilization of a training dataset, which is used as the measure domain, to analyze any similar data, which is used as the analyzed domain. Furthermore, instead of using collected data as the analyzed domain, it can be designed as a dictionary or as a grid of representative data points that capture the essential structure of the MGC diffusion.

In this paper, we present a data discretization methodology for analyzing a contiguous domain also called the analyzed domain. The presented discretization is obtained by constructing a uniform grid that will serve as a discrete version of the analyzed domain. This discretization is designed to approximate

the continuous MGC diffusion process by a discrete random walk process. More precisely, the resolution of the uniform grid is related (via an upper bound) to the approximation quality of the steady-state stationary distribution of the MGC diffusion by the discrete stationary distribution of the random walk. Therefore, the size of the constructed grid, which is based on this resolution, is not determined by the size of the analyzed data[1], but rather by the properties of the underlying MGC diffusion process and its stationary distribution.

The utilization of two separated sets of multidimensional points introduces a different approach for the analysis of very large problems. Big Data is evolving and the size of data becomes bigger every day. Twitter has more than 250 million tweets per day, Google has more than 1 billion queries per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day [23]. The data produced nowadays is estimated in the order of zettabytes (10^{10}). Current big data analytics methods focus on distributed and parallel methods such as MapReduce [6] or Hadoop [25]. However, these methods are not always the best analytics tool [16]. Hence, there is a need to find better and more efficient analysis techniques. Distribution based analysis reduces the task of data analytics to the determination of the relation between each multidimensional data point and the distribution. Any data analysis task that can be reduced to this relation (such as clustering, anomaly detection and classification) can be processed in a computational complexity that is unrelated to the size of the data. The resulting computational complexity in this case is dictated by the preprocessing density estimation step and by the kernel discretization. Both steps mostly depend on the geometry of the distribution and not on the number of data points.

The achieved DM embeddings in both the original version [4, 15] and the MGC version [2, 3], are obtained by the principal eigenvectors of the corresponding diffusion operator. These eigenvectors represent the long-term behavior of the diffusion process and capture its metastable states [12] as it converges to the unique stationary distribution. The properties of the underlying DM diffusion process were related to the density of the data in [17, 18]. Specifically, the potential of the diffusion process, which determines its stationary distribution, was shown to correlate with an underlying data potential that either generates the data densities or it is defined by them. Furthermore, this underlying potential was shown to encapsulate clustering patterns in the data. In this paper, we show similar results about the relation between the MGC diffusion potential and the underlying data potential. In addition, we examine the relations between the stationary distribution of the MGC diffusion and the data distribution as expressed by its densities. We show that the stationary distribution is a low-pass filtering version of the input data densities, where the filter bandwidth corresponds to the geometric properties of the MGC kernel. Finally, this band-

[1]The grid size is determined by neither the size of the analyzed domain nor by the size of the measure domain, but merely by the analytic properties of the kernel and the underlying measure.

width is also related to the required resolution of the obtained grid based on the approximation quality upper bound. Therefore, highly oscillatory densities will require a wide filter that results in a dense grid. More regular densities, on the other hand, will allow for the use of a narrower filter that results in a sparser grid.

This paper considers the case of a contiguous analyzed domain. An embedding of such a domain into a low-dimensional space should be preceded by a discretization process of both the analyzed domain and the corresponding MGC based operator (i.e., to obtain a kernel matrix.) Similar discretizations of a diffusion process, based on its stable (steady) and metastable states, were discussed in [12, 11, 22, 13, 7, 20]. Common approaches for this purpose invoke the Galerkin (also known as Petrov-Galerkin) discretization method [9, 24] that results in a finite partition of the diffusion domain. We use a representative set, which is sampled from such a partition, to construct a finite-rank diffusion operator. The stationary distribution of the resulting process is shown to approximate, up to a controllable error, the stationary distribution of the original continuous diffusion process. Moreover, based on the spectral analysis of this finite-rank diffusion operator, we provide a constructive algorithm that approximates the measure-based diffusion maps of the contiguous domain.

The paper is organized as follows: Section 2 presents the problem setup and discusses several previous results concerning data distributions underlying potentials and stationary distributions in the context of stochastic differential equations (SDE). The original and measure-based DM frameworks are discussed in Section 3. This section also explores the relations between data distribution, the diffusion potential and the resulting stationary distribution. The main results of this paper are established in Section 4, which presents the data discretization process that analyzes the resulting associated diffusion operator. Section 5 introduces a constructive algorithm for computing the discretized diffusion maps of the analyzed domain. The proposed methodology is demonstrated in Section 6 using analytically defined data potentials.

2. Problem Setup

Consider a big dataset $X \subseteq \mathbb{R}^m$, such that for any practical purpose the size of the dataset can be considered infinite. In this case, it is impractical, if not impossible, to consider or iterate over individual data points in X. Instead, we suggest to represent the dataset via the density of data points in it, which is modeled by a suitable density function $q : \mathbb{R}^m \to \mathbb{R}$ and a corresponding probability measure $d\mu(x) = q(x)dx$. Thus, high-probability areas of μ represent dense regions where data is clustered in X, whereas low-probability areas represent sparse regions, which separate between clusters and data points in these areas, are considered anomalous.

Let $\Gamma \subseteq \mathbb{R}^m$ be the measure domain i.e., the support of the density function

q. Assume that the analyzed dataset X is dense[2] in a contiguous domain Ω. Thus, the exact cardinality of X and the exact positions of its members in Ω are insignificant. In this paper, we focus on analyzing the contiguous dataset Ω, namely the analyzed domain, given the measure μ, instead of directly analyzing the dataset X. Therefore, we use the term measure domain for both X and the measure space (Γ, μ) interchangeably, and we assume that these two representations of the data are indistinguishable from the analysis perspective.

There are two possible scenarios for the origin of the data analysis settings in this paper, as described in [17, 18]:

1. The data is sampled from a dynamical system in equilibrium. We assume that the system is defined at time t by the SDE $\dot{x} = -\nabla U(x) + \sqrt{2}\dot{\omega}$ where the dot on a variable means differentiation with respect to time, U is the free energy at x (also referred to as the potential at x) and $\omega(t)$ is an n-dimensional Brownian motion process. In this case, there is an explicit notion of time and the transition probability densities of the system satisfy the forward and backward Fokker-Plank equation. The probability distribution in this case is defined by $p(x) = \exp(-U(x))$ (normalized to have a unit L^1 norm).
2. The data is randomly sampled from some probability distribution $p(x)$ without knowledge of an underlying SDE. In this case, we define the potential $U(x) = -\log p(x)$ that is based on the observed distribution of the sampled data.

Further details concerning this subject are given in [17, 18].

In both scenarios, the steady state probability density is identical and it satisfies $p(x) = \exp\big(-U(x)\big)$ and $U(x) = -\log p(x)$. Therefore, when the possible time dependency in the data is not directly considered, only the features of the underlying potential $U(x)$ and the probability distribution come into play. Thus, for the purpose of data analysis tasks, such as clustering, classification and anomaly detection, the properties of the underlying data potential are crucial and in many cases sufficient for the analysis. For example, many clustering patterns in Big Data originate from the topography of such underlying potentials. In this paper, we present an embedding method that preserves these properties while discretizing the analyzed dataset to a finite grid and thus reducing the dimensionality of the data.

3. Measure-based Diffusion Maps

The embedding method in this paper utilizes the measure-based diffusion embedding scheme that is presented in [2], which in turn enhances the Diffusion Maps (DM) framework from [4]. This section provides a brief overview of this diffusion-based embedding method. Both methods embed the data using a Markovian diffusion process that captures the relations between data points

[2]By saying that X is dense in Ω we mean that X is a ρ-net in Ω, with $0 < \rho \ll 1$.

and reveals the underlying data patterns. The grid construction in this paper is aimed to preserve the nature of the underlying data patterns by approximating its potential and by approximating the resulting stationary distribution of the underlying measure-based diffusion process. The properties of these potential and stationary distribution, as well as the relations between them, are discussed in Section 3.3.

3.1. Diffusion Maps

The DM framework is based on constructing a Markovian (random-walk) diffusion process that follows the geometry of the analyzed dataset X and reveals dominant patterns in it [4]. This process is defined by a set of affinities $k : X \times X \to \mathbb{R}$ that represent local relations (e.g., similarities or distances) between neighboring data points. The affinity kernel k is normalized by a set of degrees $\nu : X \to \mathbb{R}$, $\nu(x) \triangleq \int_X k(x,y)d\mu(y)$ to obtain transition probabilities $p : X \times X \to \mathbb{R}$, $p(x,y) \triangleq k(x,y)/\nu(x)$ between data points and define a random-walk process based on these probabilities. Here, μ represents the density of X. Under mild conditions on the kernel k, the resulting transition probability operator has a discrete decaying spectrum of eigenvalues $1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, 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, $\psi_t : X \to \mathbb{R}^{\eta_t}$,

$$\psi_t(x) = (\lambda_1^t \phi_1(x), \lambda_2^t \phi_2(x), \ldots, \lambda_{\eta_t}^t \phi_{\eta_t}(x)), \tag{3.1}$$

where η_t is the numerical rank of the power t of the associated integral operator. The obtained embedded representation of the data expresses the diffusion distances as Euclidean distances and the underlying patterns in the data such as clusters and differences between normal and abnormal regions.

Usually, the Gaussian affinities $k(x,y) = \exp(-\|x-y\|^2/2\varepsilon)$, for some suitable $\varepsilon > 0$, are used for the construction of the diffusion map. When the dataset X is sampled from a low dimensional manifold, its tangential spaces can be utilized to express the infinitesimal generator of the diffusion transition operator and its symmetric conjugate, which is usually referred to as the diffusion affinity kernel in terms of the Laplacian operators on the manifold. A connection between the constructed diffusion process and the underlying potential $U(x)$ is given in [18, 17]. It is shown there that the DM diffusion process operates according to Fokker-Plank equation with the potential $2U(x)$. In addition, it is shown that the degrees $\nu(x)$ approximate the probability distribution $q(x)$. When $\varepsilon \to 0$, we have $\nu(x) = q(x) + \frac{\varepsilon}{2}\Delta q(x) + O(\varepsilon^{3/2})$.

3.2. Measure-based Gaussian correlation (MGC) kernel

Recall that $\Gamma \subset \mathbb{R}^m$ is the measure domain i.e., the support of $d\mu(r) = q(r)dr$ satisfies $\int_\Gamma d\mu(r) = \int_{\mathbb{R}^m} d\mu(r) = 1$. For the rest of this paper, integrals with unspecified domains are taken over the whole space $\mathbb{R}^m$. The MGC kernel [2] function $k_\varepsilon : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is defined as

$$k_\varepsilon(x,y) \triangleq \int g_\varepsilon(x-r) g_\varepsilon(y-r) d\mu(r), \tag{3.2}$$

or equivalently,

$$k_\varepsilon(x,y) = g_{2\varepsilon}(x-y)\int g_{\varepsilon/2}\left(\frac{x+y}{2}-r\right)d\mu(r), \tag{3.3}$$

where $g_\varepsilon : \mathbb{R}^m \to \mathbb{R}$ is defined[3] as

$$g_\varepsilon(t) \triangleq \frac{1}{(\pi\varepsilon)^{m/2}} e^{-\|t\|^2/\varepsilon}. \tag{3.4}$$

It should be noticed that although the analyzed domain Ω is a subset of $\mathbb{R}^m$, the domain of k_ε and the consequent mathematical elements are defined over $\mathbb{R}^m$.

Normalization of the MGC kernel by the degrees function

$$\nu_\varepsilon(x) \triangleq \int k_\varepsilon(x,y)dy, \tag{3.5}$$

yields the associated diffusion transition probabilities kernel

$$p_\varepsilon(x,y) = k_\varepsilon(x,y)/\nu_\varepsilon(x). \tag{3.6}$$

The associated diffusion operator $P_\varepsilon : L^2(\mathbb{R}^m) \to L^2(\mathbb{R}^m)$ is

$$P_\varepsilon f(x) \triangleq \int p_\varepsilon(x,y)f(y)dy. \tag{3.7}$$

As was proved in [2], P_ε is a compact, positive definite operator whose L^2 norm is one. Therefore, it has a discrete spectrum that decays from one to zero, and it can be utilized in the diffusion maps framework. Moreover, it was proved in [2] that the infinitesimal generator of P_ε is of the form Laplacian + potential. Therefore, the associated Markov process $\{X_t\}_t$ is indeed a diffusion process. The relation between this potential of the diffusion process and the original data potential is discussed in Section 3.3.

3.3. Underlying potential and stationary distribution

The transfer probability operator P_ε (Eq. 3.7) defines an ergodic Markovian diffusion process over $\mathbb{R}^m$ [2]. Therefore, this process converges to a steady-state stationary distribution as it advances through time. In this section, we analyze this stationary distribution and relate it to the initial distribution of the measure domain as represented by the measure μ.

We consider two cases that are based on the value of the meta-parameter ε. The first case, which is discussed in Section 3.3.1, is when $\varepsilon > 0$. In this case, P_ε represents a discrete-time random walk process. The stationary distribution is

[3] In fact in [2] a slightly different definition was used for $g_\varepsilon(t) = e^{-\|t\|^2/\varepsilon}$. As shown in this section, the current normalization results in interesting interpretation of the stationary distribution while the results from [2] remain valid.

obtained when the number of time-steps tends to infinity, and it is shown to be a regularized version (via an application of a low-pass filter) of the initial data distribution. The second considered case, which is discussed in Section 3.3.2, is when $\varepsilon \to 0$. In this case, the underlying diffusion process becomes continues and can be analyzed via the infinitesimal generator of P_ε. We present a relation between this continuous-time process and the underlying data potential and show that the stationary distribution in this case converges to the data distribution μ.

Section 4 discretizes the analyzed domain Ω to provide a discrete-time and discrete-space diffusion process that is defined by a finite matrix. Then, the discretized diffusion process, which operates on a finite grid, is compared to the continuous one from this section. The relations between their steady-state distributions are analyzed.

3.3.1. Discrete-time random walks ($\varepsilon > 0$)

In this section, we analyze the Markovian process that is defined by P_ε, $\varepsilon > 0$. The transition probabilities at discrete time steps $t = 1, 2, \ldots$ are represented by the powers P_ε^t of the diffusion operator from Eq. 3.7. The meta-parameter ε serves two roles in this setting. The first role comes directly from the definition of the MGC kernel (see Eq. 3.2), which is used in Eq. 3.6 to define the transition probabilities of this process. In this context, ε quantifies the notion of spatial neighborhoods. The second role of ε is to quantize the diffusion time in order to obtain a discrete-time random walk process. Thus, big values of ε result in longer time steps and in a wider spread of the diffusion within each time step.

For a given $\varepsilon > 0$, the diffusion random-walk process, which is defined by P_ε^t, $t = 1, 2, \ldots$, is an ergodic Markov process. Therefore, it has a unique steady state stationary distribution that is achieved when $t \to \infty$. This stationary distribution represents the equilibrium state to which the process converges over time. Usually it reveals important meaningful patterns of the data [17, 18]. For example, the stationary distribution correlates with data clustering patterns, which are represented by local minima of the underlying data potential U (see Section 2).

Consider the conjugate operator $P_\varepsilon^* : L^2(\mathbb{R}^m) \to L^2(\mathbb{R}^m)$, which is defined by

$$P_\varepsilon^* f(x) \triangleq \int p_\varepsilon(y, x) f(y) dy. \tag{3.8}$$

This operator transfers probability measures over a discrete time period of ε time units. More specifically, if $X_t \sim f(x)$, then $X_{t+\varepsilon} \sim P_\varepsilon^* f(x)$. The stationary distribution of the associated diffusion process is the invariant probability measure of P_ε^*, i.e. it is the (normalized) eigenfunction of P_ε^* that corresponds to eigenvalue 1.

Obviously, the spectra of P_ε and P_ε^* are identical, therefore, the maximal eigenvalue of P_ε^* is 1. Lemma 3.1 shows that the stationary distribution is the degrees function $\nu_\varepsilon(x)$.

Lemma 3.1. *Let P_ε be the operator defined in Eq. 3.7, then the stationary distribution of the associated diffusion process is the degree function $\nu_\varepsilon(x)$ defined in Eq. 3.5.*

Proof. In order to prove the lemma we have to show that $P_\varepsilon^* \nu_\varepsilon = \nu_\varepsilon$. Indeed, $P_\varepsilon^* \nu_\varepsilon(x) = \int p_\varepsilon(y,x)\nu_\varepsilon(y)dy = \int \frac{k_\varepsilon(x,y)}{\nu_\varepsilon(y)}\nu_\varepsilon(y)dy = \nu_\varepsilon(x)$. □

Proposition 3.2, whose proof is given in Appendix A.1, shows that the stationary distribution ν_ε is a convolution of the density function $q(r)$ with g_ε, which constitutes a probability measure over $\mathbb{R}^m$.

Proposition 3.2. *The stationary distribution $\nu_\varepsilon(x) = q \star g_\varepsilon(x)$ of the diffusion process, which is defined by P_ε, is a smoothed version of the underlying measure $d\mu(r) = q(r)dr$, where $\star$ is the convolution operator and $g_\varepsilon(x)$ is defined in Eq. 3.4. Furthermore, its L^1 norm in $\mathbb{R}^m$ is $\|\nu_\varepsilon\|_{L^1(\mathbb{R}^m)} = 1$.*

As Proposition 3.2 shows, the stationary distribution ν_ε corresponds to a convolution operation of the density function $q(r)$ with a Gaussian function. This operation acts as a multi-dimensional low-pass filter on q. Therefore, the resulted stationary distribution satisfies $\hat{\nu_\varepsilon}(\omega) = \hat{q}(\omega) \cdot e^{-\varepsilon\|\omega\|^2/4}$ in the Fourier domain. This property introduces a third role for ε to be a frequency threshold. Thus, high frequencies in respect to ε are damped, where low frequencies remain unaffected. In other words, this operation restricts the oscillatoriness of q based on the threshold ε. An effective choice of ε should filter out the high frequencies that arise from undesirable phenomena such as noise, while preserving significant frequencies, which are required for pattern recognition (e.g., distinguishing between data clusters and anomaly).

3.3.2. Infinitesimal case ($\varepsilon \to 0$)

In Section 3.3.1, the value of ε was related to the frequency structure of the data density. In practical applications, this density originates from a finite (albeit massively big) dataset. Thus, the spectrum of such a density is bounded from above, therefore there exists a minimal meaningful value for the frequency threshold ε. This observation also correlates with the interpretation of ε that determines the spatial neighborhood size. Indeed, when dealing with discrete data, neighborhoods, which are too small, will be empty (except for their original centers) that will result in a degenerate diffusion process. However, in the theoretical settings of continuous data, the value of ε can be chosen to be arbitrarily small. Then, we can consider the case of $\varepsilon \to 0$, which is analyzed in this section.

Proposition 3.2 presented a relation between the stationary distribution of the underlying MGC diffusion process with a given $\varepsilon > 0$ and the initial density of the data. Corollary 3.3 extends this result to $\varepsilon \to 0$. This corollary shows that as the diffusion process becomes continuous in time, its steady-state stationary distribution converges to the data density distribution. Therefore, from any initial condition, this process will stabilize over time to reveal the underlying distribution of the analyzed data.

Corollary 3.3. *In the infinitesimal case of $\varepsilon \to 0$, the stationary distribution from Proposition 3.2 converges to* $\lim_{\varepsilon \to 0} \nu_\varepsilon(x) = q(x)$.

Proof. Due to the definition of g_ε (Eq. 3.4), $\lim_{\varepsilon \to 0} g_\varepsilon(x) = \delta(x)$, where $\delta(x)$ is the Dirac delta function. Therefore, according to Proposition 3.2, $\lim_{\varepsilon \to 0} \nu_\varepsilon(x) = q(x)$. □

The infinitesimal behavior of the original DM diffusion process with $\varepsilon \to 0$ in [4, 15, 17, 18] was analyzed via the infinitesimal generator of its transition probabilities operator. Specifically, it is shown in [4, 15] that under proper normalization of the sampling density (i.e., using an anisotropic kernel with $\alpha = 1/2$), this infinitesimal generator converges to the form of Laplacian + potential. This result was related to the Fokker-Plank equations in [17, 18]. There, it is shown that the backward transition operator of the isotropic DM diffusion process corresponds to the backward Fokker-Plank SDE with potential $2U$ (where U is the underlying potential of the data), and the anisotropic version (with $\alpha = 1/2$) corresponds to the Fokker-Plank SDE with potential U. Proposition 3.4 shows similar results for the MGC diffusion process, without requiring to apply anisotropic normalizations. The proof of this theorem is based on the analysis of the MGC infinitesimal generator that appears in Theorem 2.36 in [2].

Proposition 3.4. *The MGC diffusion process, which is defined by the backward transition operator P_ε, is associated with the backward Fokker-Plank equation $\dot{f}(x,t) = -\Delta(U(x)) + \sqrt{2}\dot{\omega}(x,t)$ (where $\omega(x,t)$ is a standard Brownian process) via its backward infinitesimal generator $\mathcal{H}_b \triangleq \lim_{\varepsilon \to 0} \frac{I - P_\varepsilon}{\varepsilon}$, which takes the form $\mathcal{H}_b f \propto \Delta f - \langle \nabla U, \nabla f \rangle$.*

Proof. Let $\mathcal{H}_b = \lim_{\varepsilon \to 0} \frac{I - P_\varepsilon}{\varepsilon}$ be the backward infinitesimal generator of the MGC diffusion process that is defined by P_ε (Eq. 3.7). Theorem 2.36 in [2] shows that this backward infinitesimal generator satisfies

$$\mathcal{H}_b f = -\frac{m_2}{m_0}\left(\Delta f + \left\langle \frac{\nabla q}{q}, \nabla f \right\rangle\right), \tag{3.9}$$

where $m_0 \triangleq \int g_\varepsilon(t)dt = (\pi\varepsilon)^{m/2}$ and $m_2 \triangleq \int g_\varepsilon(t)(t^{(j)})^2 dt$ for some coordinate j. According to the discussion in Section 2, $U(x) \triangleq -\ln(q(x))$, then by performing the appropriate substitution in Eq. 3.9 we get the form in the proposition for $\mathcal{H}_b$. This result is similar to the one obtained for the anisotropic DM diffusion process in [17, 18]. Similarly to the result there, the associated stochastic process in our case is defined via the following backward Fokker-Planck equation[4]

$$\dot{f}(x,t) = -\nabla(U(x)) + \sqrt{2}\dot{\omega}(x,t),$$

as stated in the proposition. □

[4]Note that if we take $\sqrt{q(x)}$ instead of $q(x)$ we get the backward Fokker-Plank SDE with $2U$ similar to the result from the isotropic DM diffusion [17, 18].

Corollary 3.3 together with Proposition 3.4 show that in the infinitesimal case $\varepsilon \to 0$, the continuous-time MGC diffusion process follows the underlying data potential and its related data distribution. Furthermore, this result does not assume any anisotropic normalization of the MGC kernel or of the diffusion transition operator. This normalization is not required since both operators inherently consider the distribution of the data and take it into account when determining the diffusion connectivity (i.e., transition probabilities of paths) between data regions. While the rest of the paper focuses on the discrete-time case when $\varepsilon > 0$, the convergence of these properties as proved in this section shows that for appropriately small ε, the diffusion process should follow the meaningful dominant trends in the data.

4. Data Discretization by Sampling the MGC Diffusion Process

In this section, we present a discretized version of the stochastic operator P_ε, denoted by $\tilde{P}_\varepsilon$. For this purpose, the analyzed domain Ω, which is assumed to be both bounded and Lebesgue-measurable, is divided into a finite number of uniformly bounded subsets. Then, the induced diffusion process over these subsets is modeled by a finite-rank operator $\bar{P}_\varepsilon$. In this section, we aim to formulate $\tilde{P}_\varepsilon$ as an approximation to the operator $\bar{P}_\varepsilon$. We provide a criterion for the above subdivision, which guarantees that the associated discrete stationary distribution approximates well the continuous stationary distribution.

4.1. Mathematical tools for data discretization

Let $\{\Omega_i\}_{i=1}^n$ be a partition of Ω into nonzero Lebesgue measure sets, which are almost surly disjoint, i.e. $\Omega = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_n$, $\ell_i \triangleq \ell(\Omega_i) \neq 0$ and $\ell(\Omega_i \cap \Omega_j) = 0$, where $\ell(A)$ is the Lebesgue measure of $A \subset \Omega$. Thus, $\sum_{i=1}^n \ell_i = \ell(\Omega)$. We define κ to be the partitioning parameter, i.e. $\kappa \triangleq \max_{i=1,\ldots,n} \operatorname{diam}(\Omega_i)$, where $\operatorname{diam}(\Omega_i) = \sup_{x,y\in\Omega_i}\{\|x-y\|\}$. Let $\chi_i(x)$ be the characteristic function of Ω_i and $h_i(x)$ be its normalized version in $L^2(\mathbb{R}^m)$ under the standard inner product in $L^2(\mathbb{R}^m)$, $\langle f, g\rangle \triangleq \int f(x)g(x)dx$, i.e. $h_i(x) \triangleq \chi_i(x)/\sqrt{\ell_i}$. Let H be the subspace of $L^2(\mathbb{R}^m)$ spanned by the orthonormal set $\{h_1, \ldots, h_n\}$ and let $\Pi : L^2(\mathbb{R}^m) \to H$ be the orthogonal projection on H, i.e. $\Pi f(x) \triangleq \sum_{i=1}^n \langle f, h_i\rangle\, h_i(x)$. Notations with double subindices are referred to similar elements defined on $\Omega \times \Omega$: $h_{ij}(x,y) = h_i(x) \times h_j(y)$, $\Omega_{ij} = \Omega_i \times \Omega_j$. We denote by $\mathbb{H}$ the subspace of $L^2(\Omega \times \Omega)$ spanned by the orthonormal set $\{h_{ij}\}_{i,j=1}^n$. Obviously, $\mathbb{H} = H \times H$. The notation $\langle\cdot,\cdot\rangle$ is referred to the inner product either in $L^2(\mathbb{R}^m)$ or in $L^2(\mathbb{R}^m \times \mathbb{R}^m)$, depending on the context. The same is true for the orthogonal projection Π.

4.2. The averaged operator $\bar{P}_\varepsilon$

We introduce the averaged version of the stochastic operator P_ε, which was defined in Eq. 3.7. Given a partition of Ω, as described in Section 4.1, we use the

Galerkin discretization to define the associated averaged kernel $\bar{k}_\varepsilon : \Omega \times \Omega \to \mathbb{R}^m$ as

$$\bar{k}_\varepsilon(x,y) \triangleq \Pi k_\varepsilon(x,y). \tag{4.1}$$

Thus, due to the definition of the projector Π, for any $(x,y) \in \Omega_i \times \Omega_j$ we have $\bar{k}_\varepsilon(x,y) = (\ell_i \ell_j)^{-1/2} \iint_{\Omega_i \times \Omega_j} k_\varepsilon(u,v) du dv$ which, up to a normalization factor, is the sum of the affinities among all pairs of data points from $\Omega_i \times \Omega_j$. Therefore, this quantity measures the affinity between Ω_i and Ω_j. The associated degree function $\bar{\nu}_\varepsilon : \Omega \to \mathbb{R}^m$ is

$$\bar{\nu}_\varepsilon(x) \triangleq \int \bar{k}(x,y) dy. \tag{4.2}$$

The averaged diffusion kernel $\bar{p}_\varepsilon(x,y) : \Omega \times \Omega \to \mathbb{R}$ and the associated averaged diffusion operator $\bar{P}_\varepsilon : L^2(\Omega) \to L^2(H)$ are defined as

$$\bar{p}_\varepsilon(x,y) \triangleq \frac{\bar{k}_\varepsilon(x,y)}{\bar{\nu}_\varepsilon(x)} \tag{4.3}$$

and

$$\bar{P}_\varepsilon f(x) \triangleq \int \bar{p}_\varepsilon(x,y) f(y) dy, \tag{4.4}$$

respectively.

Proposition 4.1, whose proof appears in Appendix A.2, shows that $\bar{P}_\varepsilon$ is a stochastic operator, whose spectral properties enables its utilization in the DM framework.

Proposition 4.1. *The averaged diffusion operator $\bar{P}_\varepsilon$ is a stochastic positive definite operator of a unit operator norm.*

Lemma 4.2, whose proof appears in Appendix A.3, shows the construction of the representative matrix of $\bar{P}_\varepsilon$ and the degrees function $\bar{\nu}_\varepsilon$ in the characteristic basis $\{\chi_1, \ldots, \chi_n\}$. This construction utilizes the Lebesgue measures $\ell_1, \ldots, \ell_n$ of the partitioning subsets $\Omega_1, \ldots, \Omega_n$, respectively, and the average value of $p_\varepsilon(x,y)$ on $\Omega_i \times \Omega_j$

$$m_{ij} \triangleq \frac{1}{\ell_i \ell_j} \iint k_\varepsilon(u,v) \chi_i(u) \chi_j(v) du dv. \tag{4.5}$$

Lemma 4.2. *The degrees function of $\bar{P}_\varepsilon$ is*

$$\bar{\nu}_\varepsilon(x) = \sum_{i,j=1}^{n} \ell_j m_{ij} \chi_i(x),$$

and the representative matrix of $\bar{P}_\varepsilon$ in the characteristic basis $\{\chi_1, \ldots, \chi_n\}$ is

$$[\bar{P}_\varepsilon]_{ij} = \frac{\ell_j m_{ij}}{\sum_{k=1}^{n} \ell_k m_{ik}},\ i,j = 1, \ldots, n.$$

Theoretically, the representative elements from Lemma 4.2 are sufficient for the construction of the averaged diffusion operator $\bar{P}_\varepsilon$ and are sufficient for computing the error between the original stationary distribution ν_ε and the averaged stationary distribution $\bar{\nu}_\varepsilon$. Alas, the average values from Eq. 4.5 are unknown. Therefore, an approximation of $\bar{P}_\varepsilon$ is presented in Section 4.3.

4.3. Sampled operator $\tilde{P}_\varepsilon$

In this section, we construct a sampled diffusion operator $\tilde{P}_\varepsilon$, which is a computable version of $\bar{P}_\varepsilon$. In Section 4.4, we provide a criterion for the partitioning of Ω and the resulting construction of $\tilde{P}_\varepsilon$, which guarantees a controllable error between the continuous stationary distribution ν_ε and the degrees function $\tilde{\nu_\varepsilon}$ of the sampled operator $\tilde{P}_\varepsilon$.

The construction of the sampled operator is similar to the construction of the averaged operator $\bar{P}_\varepsilon$, but instead of using the averaged values m_{ij}, as described in Section 4.2, we evaluate the kernel on a sampled dataset. Specifically, we sample n representatives $x_1, \ldots, x_n$ from Ω, one from each cell, i.e., $x_i \in \Omega_i$, $i = 1, \ldots, n$. For such a set, we define the piecewise constant kernel $\tilde{k}_\varepsilon : \Omega \times \Omega \to \mathbb{R}$ to be

$$\tilde{k}_\varepsilon(x, y) \triangleq k_{ij},\ x \in \Omega_i,\ y \in \Omega_j,$$

where

$$k_{ij} = k_\varepsilon(x_i, x_j). \tag{4.6}$$

Similarly to the definitions from Eqs. 4.2, 4.3 and 4.4, the sample degree function $\tilde{\nu_\varepsilon} : \Omega \to \mathbb{R}$ is defined by

$$\tilde{\nu_\varepsilon}(x) \triangleq \int \tilde{k}_\varepsilon(x, y)dy, \tag{4.7}$$

the sampled diffusion operator $\tilde{p}_\varepsilon : \Omega \times \Omega \to \mathbb{R}$ is defined as

$$\tilde{p}_\varepsilon(x, y) \triangleq \frac{\tilde{k}_\varepsilon(x, y)}{\tilde{\nu_\varepsilon}(x)} \tag{4.8}$$

and the associated stochastic operator $\tilde{P}_\varepsilon : L^2(\Omega) \to L^2(H)$ is defined by

$$\tilde{P}_\varepsilon f(x) \triangleq \int \tilde{p}_\varepsilon(x, y)f(y)dy. \tag{4.9}$$

Thus, similarly to the results in Lemma 4.2, we get

$$\tilde{\nu_\varepsilon}(x) = \sum_{i,j=1}^{n} \ell_j k_{ij} \chi_i(x), \tag{4.10}$$

and the representative matrix of $\tilde{P}_\varepsilon$ in the characteristic basis $\{\chi_1, \ldots, \chi_n\}$ is

$$[\tilde{P}_\varepsilon]_{ij} = \frac{\ell_j k_{ij}}{\sum_{s=1}^{n} \ell_s k_{is}},\ i, j = 1, \ldots, n. \tag{4.11}$$

Thus, if the partition is uniform such that $\ell_i = \ell_j$ for any $i, j = 1, \ldots, n$, then $[\tilde{P}_\varepsilon]_{ij} = k_{ij} / \sum_{s=1}^{n} k_{is}$. Obviously, if $\{\lambda_k\}_{k=1}^n$ and $\{\phi_k\}_{k=1}^n$ are the eigenvalues and eigenvectors, respectively, of $[\tilde{P}_\varepsilon]$, then the eigenvectors of $\tilde{P}_\varepsilon$ are $\Phi_k : \Omega \to \mathbb{R}$, defined by

$$\Phi_k(x) \triangleq \sum_{i=1}^{n} \phi_{ki} \chi_i(x),\ k = 1, \ldots, n, \tag{4.12}$$

where ϕ_{ki} denotes the i-th coordinate of ϕ_k. The corresponding eigenvalues are $\{\lambda_k\}_{k=1}^n$.

4.4. Discretization criterion of Ω

Due to the significance of the stationary distribution of the continuous diffusion process, which was explained in details in Section 3.3, our discretization criterion is based on the approximation of the continuous stationary distribution ν_ε by the sampled degrees function $\tilde{\nu}_\varepsilon$. More specifically, given a predefined positive parameter δ, our goal is to discretize Ω by uniform sampling and, accordingly, to construct the sampled operator $\tilde{P}_\varepsilon$, such that the resulted degrees function $\tilde{\nu}_\varepsilon$ provides $\|\tilde{\nu}_\varepsilon - \nu_\varepsilon\|_\infty \leq \delta$. The main result of this section is formulated in Theorem 4.5, which quantifies $\|\tilde{\nu}_\varepsilon - \nu_\varepsilon\|_\infty$ as a function of the partitioning parameter κ, the width of the Gaussian ε, and the Lipschitz constant of q.

Lemma 4.3, whose proof is given in Appendix A.4, provides a bound for the difference between m_{kl}, the averaged value of the kernel k_ε on $\Omega_k \times \Omega_l$, and the value $k_\varepsilon(x_k, x_l)$, evaluated for an arbitrary pair $(x_k, x_l) \in \Omega_k \times \Omega_l$. For its statement, we define the quantity $b_{a,\varepsilon,\kappa}$ which depends on the Lipschitz parameter a of the density q, the Gaussian's width ε and the partitioning parameter κ:

$$b_{a,\varepsilon,\kappa} \triangleq (2\varepsilon\pi)^{-m/2}(a\kappa + \min\left\{1, 2\kappa(\varepsilon e)^{-1/2}\right\}(1 + a(m\varepsilon/2)^{1/2}). \tag{4.13}$$

Lemma 4.3. *If the density function q satisfies $q(x) \leq 1$ for any $x \in \mathbb{R}^d$ and its Lipschitz parameter is a, then for any arbitrary choice of pairs $(x_k, x_l) \in \Omega_k \times \Omega_l$ and $(\xi_k, \xi_l) \in \Omega_k \times \Omega_l$, $|k_\varepsilon(\xi_k, \xi_l) - k_\varepsilon(x_k, x_l)| \leq b_{a,\varepsilon,\kappa}$.*

For a very oscillatory density function q, whose Lipschitz constant a is large, the bound in Lemma 4.3 may be too crude. For such cases, Lemma 4.4, which does not use the Lipschitzness of q, should be considered. For its statement, we define the following quantity that depends on the Gaussian's width ε and on the partitioning parameter κ, to be

$$h_{\varepsilon,\kappa} \triangleq 2\pi^{-m/2}\varepsilon^{-(m+1)/2}\kappa. \tag{4.14}$$

The proof of Lemma 4.4 appears in Appendix A.5.

Lemma 4.4. *For any arbitrary choice of pairs $(x_k, x_l) \in \Omega_k \times \Omega_l$ and $(\xi_k, \xi_l) \in \Omega_k \times \Omega_l$, $|k_\varepsilon(\xi_k, \xi_l) - k_\varepsilon(x_k, x_l)| \leq h_{\varepsilon,\kappa}$.*

Theorem 4.5 concludes the above and by establishing a partitioning criterion that guarantees a certain bound for $\ell(\Omega)^{-1}\|\nu - \tilde{\nu}\|_\infty$. For this purpose, we define the quantity $f_{a,\varepsilon,\kappa}$ which depends on the Lipschitz parameter a of the density q, the Gaussian's width ε and on the partitioning parameter κ, to be

$$f_{a,\varepsilon,\kappa} \triangleq \min\{b_{a,\varepsilon,\kappa}, h_{\varepsilon,\kappa}\}. \tag{4.15}$$

Theorem 4.5. *If the density function q satisfies $q(x) \leq 1$ for any $x \in \mathbb{R}^d$ and its Lipschitz parameter is a, then $\ell(\Omega)^{-1}\|\nu(x) - \tilde{\nu}(x)\|_\infty \leq f_{a,\varepsilon,\kappa}$.*

Proof. Using the definitions of ν_ε and $\tilde{\nu}_\varepsilon$ in Eqs. 3.5 and 4.7, respectively, for any $x \in \Omega_i$ we have

$$\begin{aligned} |\tilde{\nu}_\varepsilon(x) - \nu_\varepsilon(x)| &= \left| \sum_j \int_{\Omega_j} k_\varepsilon(x,y)dy - \sum_j \ell_j k_{ij} \right| \\ &\leq \max_{j=1,\ldots,n} \left| \ell_j^{-1} \int_{\Omega_j} k_\varepsilon(x,y)dy - k_{ij} \right| \cdot \sum_j \ell_j \\ &= \ell(\Omega) \cdot \max_{j=1,\ldots,n} \left| \ell_j^{-1} \int_{\Omega_j} k_\varepsilon(x,y)dy - k_{ij} \right|. \end{aligned}$$

According to the mean value theorem, for any $x \in \Omega_i$ there is $y_j(x) \in \Omega_j$, for which $\ell_j^{-1} \int_{\Omega_j} k_\varepsilon(x,y)dy = k_\varepsilon(x, y_j(x))$. Therefore, due to Lemmas 4.3 and 4.4, for any $x \in \Omega_i$, $\left| \ell_j^{-1} \int_{\Omega_j} k_\varepsilon(x,y)dy - k_{ij} \right| \leq f_{a,\varepsilon,\kappa}$. Thus, as a consequence, $\ell(\Omega)^{-1} \left\| \nu(x) - \tilde{\nu}(x) \right\|_\infty \leq f_{a,\varepsilon,\kappa}$. □

5. Mapping algorithm

In this section, we present a constructive algorithm that provides a family of MGC-based diffusion embeddings of a contiguous domain. Based on the analysis from Section 4, the algorithm models a discrete MGC-based diffusion process over a finite data set that is sampled from this domain. Finally, the analysis of the discrete process is generalized to the original domain up to a controllable approximation error.

Algorithm 1 is applied to a discretized version of the analyzed domain Ω, provided by its partitioning, which is done according to Theorem 4.5. Specifically, the partitioning parameter κ is chosen by the user to guarantee that the bound of the difference between the discrete and the continuous steady states $f_{a,\varepsilon,\kappa}$ from Eq. (4.15) is sufficiently tight. Then, based on a random choice of representatives from the resulting partition, a discrete finite-rank diffusion operator $[\tilde{P}_\varepsilon]$ is formulated using Eq. 4.11. This matrix represents the sampled operator P_ε from Eq. 4.9 in the characteristic bases associated with the partition. Finally, spectral decomposition of $[\tilde{P}_\varepsilon]$ provides a family of DM for the analyzed domain Ω, which depends on a time parameter t. More specifically, the time parameter t serves as a scale parameter, and the resulting spectral decomposition is used to construct multi-scale DM, as shown in Eq. 3.1. Algorithm 1 requires the knowledge of the Lipschitz constant a or its upper bound. Though such a constant can be estimated, in general finding a good estimate for a can be a formidable task. In most applications where Lipschitzian optimization techniques have been proposed, an adaptive approximation of a was proposed - see [10] and references within. This global approximation can be used as a preprocessing step for estimating a.

Algorithm 1: Discretized MGC based diffusion maps

Input: A bounded region of interest Ω, Lipschitz data density q with constant a, Gaussian width ε, diffusion transition time t, accuracy threshold δ.

Output: Discretized diffusion map of Ω, $\psi_t : \Omega \to \mathbb{R}^{\eta_t}$, $\psi_t(x) = (\lambda_1^t \Phi_1(x), \ldots, \lambda_{\eta_t}^t \Phi_{\eta_t}(x))$, where η_t is the numerical rank of $\tilde{P}_\varepsilon^t$ and $\{\lambda_j^t\}_{j=1}^n$ and $\{\Phi_j\}_{j=1}^n$ are its eigenvalues and eigenvectors, respectively. The degrees function $\tilde{\nu}_\varepsilon$ of $\tilde{P}_\varepsilon$ provides $\ell(\Omega)^{-1} \left\| \nu_\varepsilon - \tilde{\nu}_\varepsilon \right\|_\infty \leq \delta$.

1: Divide Ω into almost surely non-zero Lebesgue measurable subsets $\Omega_1, \ldots, \Omega_n$, whose Lebesgue measures are $\ell_1, \ldots, \ell_n$, respectively, where the partition parameter κ satisfies $f(a, \varepsilon, \kappa) \leq \delta$ (see Eq. 4.15).

2: From each subset Ω_j, choose (arbitrarily) a single point $x_j \in \Omega_j$.

3: Apply Eqs. 3.2, 4.6 and 4.11 to construct the $n \times n$ representative matrix $[\tilde{P}_\varepsilon]$ in the basis of the characteristic functions $\{\chi_j\}_{j=1}^n$, associated with the partition $\{\Omega_j\}_{j=1}^n$.

4: Apply an eigen-decomposition to $[\tilde{P}_\varepsilon]$ to provide its eigenvalues and eigenvectors, $\{\lambda_k\}_{k=1}^n$ and $\{\phi_k\}_{k=1}^n \subset \mathbb{R}^n$, respectively.

5: Compute the eigenfunctions $\{\Phi_k\}_{k=1}^n$ of $\tilde{P}_\varepsilon$ using Eq. 4.12.

6: Using Eq. 3.1, construct the diffusion maps $\{\psi_t\}_{t \geq 0}$, $\psi_t : \Omega \to \mathbb{R}^{\eta_t}$, $\psi_t(x) \triangleq (\lambda_1^t \Phi_1(x), \ldots, \lambda_{\eta_t}^t \Phi_{\eta_t}(x))$.

6. Experimental Results

This section presents two examples that demonstrate the principles of the MGC discretization. The first example presents an analysis of a density function for which the stationary distribution is analytically known. The discrete stationary distribution in this case is compared to the analytical stationary distribution and the resulting error is related to the bounds from Theorem 4.5. The second example describes an embedding of a 4-well potential. In this example, we demonstrate how to utilize the proposed algorithm for toy example clustering.

Figure 6.1 presents the given density function. 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,

$$q(r) = \frac{1}{5} \chi_{[0,1]\times[0,1]}(r) + \frac{4}{5} \chi_{[3,4]\times[3,4]}(r). \tag{6.1}$$

Proposition 3.2 describes the relation between the density function and the stationary distribution. Given $\varepsilon = 1$, the convolution $\nu(x) = q(r) \star g_\varepsilon(x)$ can be analytically solved as

$$\nu(x_1, x_2) = \frac{1}{5} H(0, 1, x_1, x_2) + \frac{4}{5} H(3, 4, x_1, x_2), \tag{6.2}$$

where $H(a, b, x_1, x_2)$ is a given by

$$H(a, b, x_1, x_2) = \frac{1}{4} \left(\text{erf}(b - x_1) - \text{erf}(a - x_1)\right) \left(\text{erf}(b - x_2) - \text{erf}(a - x_2)\right),$$

and $\text{erf}(x)$ is the Gauss error function.

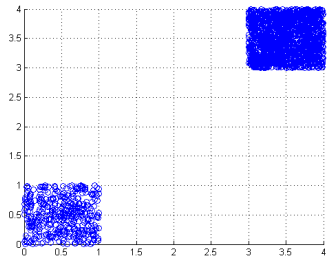


Figure 6.1: Dataset distribution

The bound in Eq. 4.15 is utilized to construct a grid over the analyzed domain, given a desired error for stationary distribution approximation, as stated in Theorem 4.5. The grid divides the space into smaller m-dimensional uniform cubes. Each grid cube edge ι relates to κ as $\iota = \kappa m^{-\frac{1}{2}}$ where κ is the partitioning parameter (see Eq. 4.14). In our example, we assume that the desired error bound ζ is given as $\zeta = 0.1$. Hence, we seek to find the maximal κ under this bound condition. The resulting κ is then translated to grid cube edge ι. Since the grid is uniform, the cube edge ι is the grid resolution. For example, in our density function given by Eq. 6.1, we have $m = 2$, $a = 0$ and $\varepsilon = 1$. Introducing $a = 0$ into $b_{a=0,\varepsilon,\kappa}$ yields that $h_{\varepsilon,\kappa} < b_{a=0,\varepsilon,\kappa}$ (see Eqs. 4.13 and 4.14). Hence, in this case, $h_{\varepsilon,\kappa}$ is a tighter bound than $b_{a=0,\varepsilon,\kappa}$ and $f_{a=0,\varepsilon,\kappa} = h_{\varepsilon,\kappa}$.

Given ζ, the value of κ can be extracted using Eq. 4.14 where $\kappa \leq \frac{\zeta}{2}\pi^{m/2}\varepsilon^{\frac{m+1}{2}}$, which, in our case, yields $\kappa \leq \frac{0.1}{2}\pi = 0.157$. Hence, the number of required grid points in each dimension is $(8/0.157)\sqrt{(2)} \approx 72$ in order to guarantee that the approximation error of the stationary distribution is less than ζ.

Table 6.1 compares between the suggested bounds and the computed actual error given the analytical stationary distribution for various values of ι . Table 6.1 suggests that the proposed bounds are not tight. The gap between the actual error and the computed bound is about 2 orders of magnitude.

ι	**Actual Error**	$h_{\varepsilon,\kappa}$	$b_{a=0,\varepsilon,\kappa}$
$\frac{1}{2}$	8.8×10^{-4}	0.45	0.54
$\frac{1}{4}$	4.4×10^{-4}	0.22	0.27
$\frac{1}{8}$	2.1×10^{-4}	0.11	0.13
$\frac{1}{16}$	0.9×10^{-4}	0.05	0.06

Table 6.1: Comparison between the actual discretization error and the estimated bound, ι is the grid resolution, the actual error is computed by $\frac{\|\nu(x)-\tilde{\nu}(x)\|_\infty}{\ell(\Omega)}$, $h_{\varepsilon,\kappa}$ and $b_{a=0,\varepsilon,\kappa}$ are the bounds from Eqs. 4.14 and 4.13, respectively

Figure 6.2 presents the analytical stationary distribution that was computed over the given density function by Eq. 6.2. Figure 6.3 complements the results in Table 6.1 by illustrating the resulting discretized stationary distribution for each configuration ι. The discretized stationary distribution was computed over a uniform grid by using Eq. 4.10, where the kernel was computed by introducing $q(r)$ from Eq. 6.1 into Eq. 3.2. Figure 6.3 shows that as the grid resolution becomes finer, the resulting discrete stationary distribution becomes smother and converges to the analytical one, which is presented in Figure 6.2. In Figure 6.3(a), the granularity in the resulted discretized stationary distribution is significant for $\iota = \frac{1}{4}$ and we can see the edges of each grid cell. For $\iota = \frac{1}{8}$, the discretized stationary distribution is smoother but the computed bound in Table 6.1 is too large. Figure 6.3(c) and Figure 6.3(d) present discretized stationary distribution with minor distortion compared to Figure 6.2.

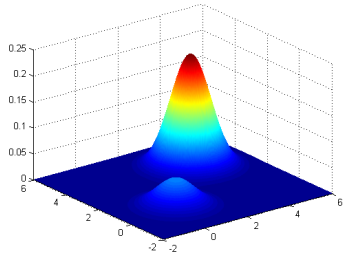
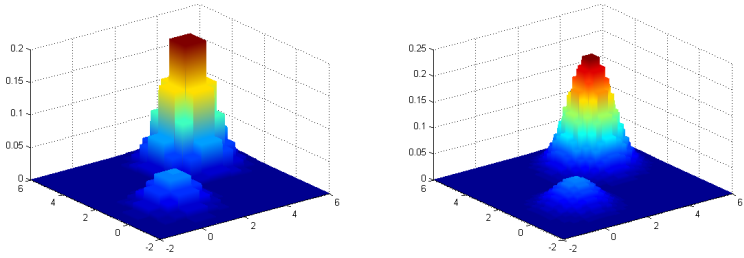
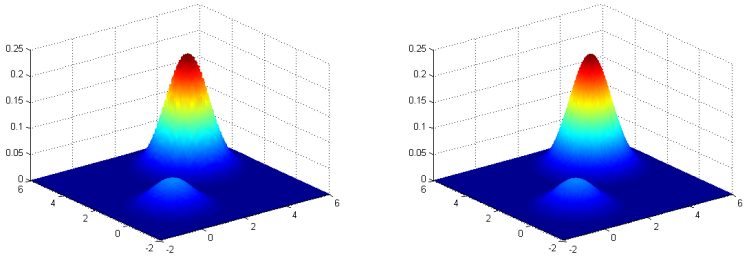


Figure 6.2: The analytical stationary distribution

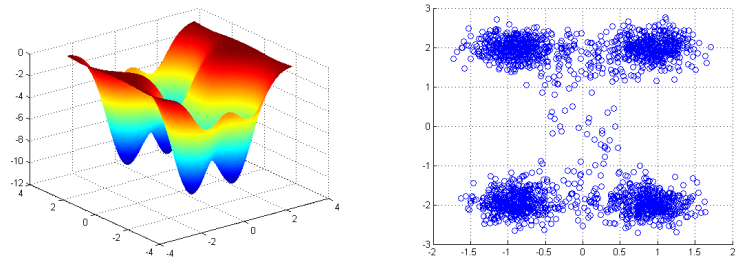


(a) $\iota = \frac{1}{4}$ (b) $\iota = \frac{1}{8}$



(c) $\iota = \frac{1}{16}$ (d) $\iota = \frac{1}{32}$

Figure 6.3: Comparing between the discrete stationary distribution and the analytical stationary distribution as a function of ι. (a) $\iota = \frac{1}{4}$. (b) $\iota = \frac{1}{8}$. (c) $\iota = \frac{1}{16}$. (d) $\iota = \frac{1}{32}$.



(a) Four-well potential (b) Dataset distribution

Figure 6.4: (a) Four-well potential. (b) Data point distribution that corresponds to the given potential

The second example in this section considers the obtained embedding based on the proposed MGC grid. First, assume that the data is sampled from a given

potential (see Fig. 6.4(a)) of the form

$$U(x,y) = -4e^{-x^2}\left(e^{-\left(y-\frac{1}{2}\right)^2} + e^{-\left(y+\frac{1}{2}\right)^2}\right) - 10\left(e^{-(y+2)^2} + e^{-(y-2)^2}\right)\left(e^{-(x-1)^2} + e^{-(x+1)^2}\right). \quad (6.3)$$

This potential consists of four local wells, which encompass most of the sampled data from the corresponding data distribution function, given by $q(x) = c \cdot \log(-U(x))$. The scaling factor c guarantees that $\int q(x)dx = 1$. In this example, 2000 data points were sampled from the described distribution. The given potential U and the resulting sampled data points are presented in Fig. 6.4. As can be seen from Fig. 6.4(b), most of the data points are concentrated around four major centers, which correspond to the four potential wells. However, there is a nonzero probability of sampling data points from areas between these centers.

In order to compute the embedding in the second example, we design a grid with 2^9 points in each dimension. The total number of grid points for the embedding is 2^{18} which is very large. The integral in Eq. 3.2 was approximated by summing over all the admissible grid points as a matrix equation $K = V^T Q V$, where $[V]_{i,j} = g_\varepsilon(x_i - r_j)$, $i = 1, \ldots, 2^{18}$, $j = 1, \ldots, 2000$, Q is a diagonal matrix with $q(r_j)$ on its j-th diagonal element and r is a set of random points drawn from the distribution $q(x)$.

The number of admissible grid points that have probabilities greater than 0.9×10^{-5} is 7636. Additionally, for the embedding we used $\varepsilon = 6$. Following the steps of Algorithm 1, the resulted matrices V and Q dimensions are $V \in \mathbb{R}^{2000\times 7636}$ and $Q \in \mathbb{R}^{2000\times 2000}$.

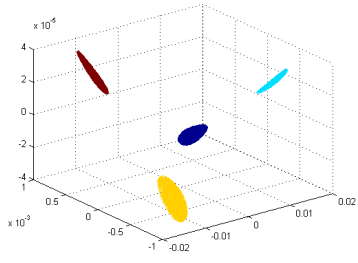


Figure 6.5: The embedding of data points sampled from a four-well potential

The embedding of the generated data points from the given potential is presented in Figure 6.5. The embedding preserves the main structure of the given potential and shows a concentration of all the data points into the corresponding clusters in the embedded space.

7. Conclusions

The MGC diffusion-based embedding methodology was presented in [2] for continuous data using a measure-based diffusion operator. This methodology is enhanced in this paper by providing a discretization scheme for its application using a finite random-walk transition matrix. The presented discretization criterion is based on approximating the steady state stationary distribution of the measure-based diffusion process from [2]. This discretization method constructs a uniform grid such that an MGC-based Markovian process defined on it approximates the stationary distribution of the underlying MGC diffusion process on the entire input data.

The MGC diffusion embedding is achieved by considering the principal eigenvectors of the MGC diffusion kernel. These eigenvectors represent the long term metastable states of the diffusion as it converges to the discussed stationary distribution. Clustered high-density areas in the resulting embedded space correspond to local minima in the underlying potential of the underlying MGC diffusion process and they can be equivalently characterized by high concentration patterns in the stationary distribution of this process. Therefore, by approximating this distribution, the presented discretized embedding map preserves important clustering patterns in the data.

The MGC diffusion stationary distribution is related to the distribution of the analyzed data as expressed by its densities and underlying potential by a low-pass filter whose bandwidth is determined by a meta-parameter ε of the MGC diffusion. The meta-parameter ε also has a geometric meaning as the size of the geometric neighborhoods that are captured by the kernel [2]. The main result in this paper (Theorem 4.5) relates this meta-parameter to the required grid resolution via an upper bound on the stationary distribution approximation error. These results provide a relation between the sparsity of the constructed grid and the regularity of the data densities. When the data density (or, equivalently, its underlying potential as described in Section 3.3) changes slowly in space, it can be faithfully captured by wide diffusion neighborhoods (i.e., having a frequency-narrow low-pass filter with a big ε value), which will be discretized to become a sparse grid. A highly oscillatory density, on the other hand, requires the stationary distribution to capture higher frequencies, thus ε should be set to a small value, which results in small diffusion neighborhoods and in a dense grid. Finally, the grid size and the resolution do not depend on the size of the input data, but they are determined only by the regularity or variability of its densities.

7.1. Future work

Future work will focus on extending and refining the measure-based diffusion discretization scheme in three main ways:

1. **Adaptive grid construction:** The constructed grid in the presented discretization scheme is uniform, thus, its resolution is controlled by the maximum variability of the underlying data potential and density. Therefore, the same grid can be utilized even when new data is added to the

analysis or when analyzing changing data as long as the oscillatory nature of its underlying potential remains similar. However, the uniform grid also has the disadvantage by not considering locally regular areas of the data density. Thus it can be redundant in smooth areas. An adaptive grid construction can optimize the grid size by considering local discretization criteria that generates sparse concentrations of grid points in areas where the diffusion potential is regular and denser concentrations where it is more oscillatory.

2. **Geometric discretization criteria:** The discretization scheme in this paper is based on preserving (or approximating) the stationary distribution of the MGC diffusion processed that is used for data embedding. This criterion preserves density and clustering patterns in the data and in the embedded space of the MGC based diffusion map. However, it does not directly consider the geometric structure of the embedded space, which is determined by metastable states of the diffusion process rather than its stationary stable steady state. Future work will consider enhancing and extending the discretization scheme by considering the embedded space geometry in the discretization criterion. Specifically, the metric structure of the embedded space, which is determined by the MGC diffusion distances [2], will be considered as an approximation goal for the grid construction algorithm.

3. **Dictionary constructions:** While the previous two aspects deal directly with optimizing the discretization process, an alternative approach is to apply dictionary learning as a post-processing phase to refine the already-constructed grid. Therefore, using this approach, the obtained grid is regarded as a coarse discretization that serves as an input for dictionary construction algorithms. Such algorithms can refine the grid by only including important grid point (under appropriate selection criteria). However, since the construction algorithm is executed on the grid instead of the input data, its performance will not depend on the size of the input dataset but rather on the properties of the discretization.

An important case that will be considered for all the above issues is when the measure fits a known model (e.g., Gaussian Mixture Model), in which case the structure of the discretized grid can be analytically computed (e.g., using a closed form expression). This case is especially important for the last issue, since a dictionary can be computed in this case without the need to construct a-priori the entire grid.

Appendix A Technical Proofs

A.1 Proof of Proposition 3.2

Proposition 3.2. *The stationary distribution $\nu_\varepsilon(x) = q \star g_\varepsilon(x)$ of the diffusion process, which is defined by P_ε, is a smoothed version of the underlying measure*

$d\mu(r) = q(r)dr$, where $\star$ is the convolution operator and $g_\varepsilon(x)$ is defined in Eq. 3.4. Furthermore, its L^1 norm in $\mathbb{R}^m$ is $\|\nu_\varepsilon\|_{L^1(\mathbb{R}^m)} = 1$.

Proof. We observe that for any $\varepsilon > 0$, $\int g_\varepsilon(t)dt = 1$. Thus, from Eqs. 3.5 and 3.2 we have

$$\begin{aligned}
\nu_\varepsilon(x) &= \int k_\varepsilon(x,y)dy \\
&= \iint g_\varepsilon(x-r)g_\varepsilon(y-r)q(r)drdy \\
&= \int \left(\int g_\varepsilon(y-r)dy\right) g_\varepsilon(x-r)q(r)dr \\
&= \int g_\varepsilon(x-r)q(r)dr.
\end{aligned}$$

Consequently, $\int \nu_\varepsilon(x)dx = \int \left(\int g_\varepsilon(x-r)dx\right) q(r)dr = \int q(r)dr = 1$. □

A.2 Proof of Proposition 4.1

In order to prove Proposition 4.1, we first define the average value of $\bar{p}_\varepsilon$ on $\Omega_i \times \Omega_j$ to be

$$\bar{p}_{ij} \triangleq \frac{1}{\ell_i \ell_j} \int_{\Omega_i} \int_{\Omega_j} p_\varepsilon(x,y)dxdy, \tag{A.1}$$

and to prove the following Lemma:

Lemma A.2. *For an operator from $L^2(\Omega)$ to itself, $\bar{P}_\varepsilon = \Pi P_\varepsilon \Pi$.*

Proof. Let $f \in L^2(\Omega)$, then from Eq. 4.1 we get

$$\begin{aligned}
\Pi P_\varepsilon \Pi f(z) &= \sum_{j=1}^{n} \langle f, h_j \rangle \, \Pi P_\varepsilon h_j(x) \\
&= \sum_{j=1}^{n} \langle f, h_j \rangle \left[\sum_{i=1}^{n} \iint p_\varepsilon(x,y) h_j(y) h_i(x) dxdy h_i(z)\right] \\
&= \sum_{i,j=1}^{n} \bar{p}_{ij} \langle f, h_j \rangle \, h_i(z) \\
&= \int \sum_{i,j=1}^{n} \bar{p}_{ij} h_i(z) h_j(y) f(y) dy \\
&= \bar{P}_\varepsilon f(z).
\end{aligned}$$

Moreover, $\left\|\bar{P}_\varepsilon\right\| = \|\Pi P_\varepsilon \Pi\| \leq \|P_\varepsilon\| = 1$. This is due to the fact that the last inequality is due to the fact that the norm of the orthogonal projector Π is 1 and the last equality was proved in [2]. Since $\bar{P}_\varepsilon \chi_\Omega(x) = \chi_\Omega(x)$, we get $\left\|\bar{P}_\varepsilon\right\| = 1$. □

Proposition 4.1. *The averaged diffusion operator $\bar{P}_\varepsilon$ is a stochastic positive definite operator of a unit operator norm.*

Proof. Due to the definition of the kernel $\bar{p}_\varepsilon$ (see Eq. 4.3), the operator $\bar{P}_\varepsilon$ is indeed stochastic. Moreover, according to Lemma A.2, we have $\left\langle \bar{P}_\varepsilon f, f \right\rangle = \langle P_\varepsilon \Pi f, \Pi f \rangle > 0$ for any $f \neq 0$ in $L^2(\Omega)$. The last equality is due to the fact that Π, as an orthogonal projection operator it is self conjugate. The last inequality is due to the positive definiteness of P_ε, as was proved in [2]. □

A.3 Proof of Lemma 4.2

Lemma 4.2. *The degrees function of $\bar{P}_\varepsilon$ is*

$$\bar{\nu}_\varepsilon(x) = \sum_{i,j=1}^{n} \ell_j m_{ij} \chi_i(x),$$

and the representative matrix of $\bar{P}_\varepsilon$ in the characteristic basis $\{\chi_1, \ldots, \chi_n\}$ is

$$[\bar{P}_\varepsilon]_{ij} = \frac{\ell_j m_{ij}}{\sum_{k=1}^{n} \ell_k m_{ik}},\ i, j = 1, \ldots, n.$$

Proof. According to the definition of $\bar{\nu}_\varepsilon$ in Eq. 4.2, $\bar{\nu}(x) = \int \bar{k}(x,y)dy = \sum_{i,j=1}^{n} m_{ij}\chi_i(x) \int \chi_j(y)dy = \sum_{i,j=1}^{n} \ell_j m_{ij} \chi_i(x)$. Additionally, $\bar{P}\chi_k(x) = \int \sum_{i,j=1}^{n} \bar{p}_{ij}\chi_i(x)\chi_j($ $\sum_{i=1}^{n} \bar{p}_{ik}\chi_i(x)\ell_k$, therefore, due to Eqs. A.1 and 4.5, the Lemma is proved. □

A.4 Proof of Lemma 4.3

Lemma 4.3. *If the density function q satisfies $q(x) \leq 1$ for any $x \in \mathbb{R}^d$ and its Lipschitz parameter is a, then for any arbitrary choice of pairs $(x_k, x_l) \in \Omega_k \times \Omega_l$ and $(\xi_k, \xi_l) \in \Omega_k \times \Omega_l$, $|k_\varepsilon(\xi_k, \xi_l) - k_\varepsilon(x_k, x_l)| \leq b_{a,\varepsilon,\kappa}$.*

Proof. Let $\Delta_\xi = \xi_k - \xi_l$, $\Delta_x = x_k - x_l$, $m_\xi = \frac{1}{2}(\xi_k + \xi_l)$ and $m_x = \frac{1}{2}(x_k + x_l)$ then, according to Eq. 3.3

$$\begin{aligned} k_\varepsilon(\xi_k, \xi_l) - k_\varepsilon(x_k, x_l) &= g_{2\varepsilon}(\Delta_\xi) \int g_{\varepsilon/2}\left(m_\xi - r\right) q(r)dr \\ &- g_{2\varepsilon}(\Delta_x) \int g_{\varepsilon/2}\left(m_x - r\right) q(r)dr \\ &= \underbrace{g_{2\varepsilon}(\Delta_\xi) \int \left[g_{\varepsilon/2}\left(m_\xi - r\right) - g_{\varepsilon/2}\left(m_x - r\right)\right] q(r)dr}_{I} \\ &+ \underbrace{\left[g_{2\varepsilon}(\Delta_\xi) - g_{2\varepsilon}(\Delta_x)\right] \int g_{\varepsilon/2}(m_x - r) q(r)dr}_{II}. \end{aligned} \quad \text{(A.2)}$$

Since $g_{2\varepsilon}(\Delta_\xi) \leq (2\varepsilon\pi)^{-m/2}$ and $\int g_{\varepsilon/2}\left(m_\xi - r\right) q(r)dr = \int g_{\varepsilon/2}\left(m_x - r\right) q(r + m_\xi - m_x)dr$, term I can be bounded by

$$|I| \leq (2\varepsilon\pi)^{-m/2} \int g_{\varepsilon/2}(m_x - r) \left|q(r + m_\xi - m_x) - q(r)\right| dr.$$

Recall that $\|\Delta_x\| \leq \kappa$ and $\|\Delta_\xi\| \leq \kappa$, therefore also $\|m_\xi - m_x\| \leq \kappa$. Thus $|q(r + m_\xi - m_x) - q(r)| \leq a\kappa$, and we get

$$|I| \leq (2\varepsilon\pi)^{-m/2} a\kappa. \tag{A.3}$$

As for term II:

$$\begin{aligned}
\int g_{\varepsilon/2}(m_x - r)q(r)dr &= \int g_{\varepsilon/2}(r)q(r + m_x)dr \\
&\leq \int g_{\varepsilon/2}(r)\left[q(m_x) + a\|r\|\right]dr \\
&= q(m_x) + \underbrace{a\int g_{\varepsilon/2}(r)\|r\|\,dr}_{III}.
\end{aligned} \tag{A.4}$$

The integrand in term III is radial. For a fixed $\|R\|$, it equals to an $(m-1)$-dimensional hypersphere's surface $S_{m-1}(\|R\|) = \frac{2\pi^{m/2}\|R\|^{m-1}}{\Gamma(m/2)}$, weighted by $g_{\varepsilon/2}(\|r\|)$. Thus we get

$$\begin{aligned}
|III| &= (\pi\varepsilon/2)^{-m/2} a \frac{2\pi^{m/2}}{\Gamma(m/2)} \int_0^\infty e^{-2x^2/\varepsilon} x^m dx \\
&= (\pi\varepsilon/2)^{-m/2} a \frac{2\pi^{m/2}}{\Gamma(m/2)} \cdot \frac{1}{2}\Gamma\left(\frac{m+1}{2}\right)\left(\frac{\varepsilon}{2}\right)^{\frac{m+1}{2}} \\
&= a(\varepsilon/2)^{1/2}\Gamma\left(\frac{m+1}{2}\right)/\Gamma(m/2) \\
&\leq a(m\varepsilon/2)^{1/2}
\end{aligned} \tag{A.5}$$

In addition, since $|\|\Delta_\xi\| - \|\Delta_x\|| \leq \|x_k - \xi_k\| + \|x_l - \xi_l\| \leq 2\kappa$, we get

$$|g_{2\varepsilon}(\Delta_\xi) - g_{2\varepsilon}(\Delta_x)| \leq (2\varepsilon\pi)^{-m/2} \min\left\{1, 2\kappa(\varepsilon e)^{-1/2}\right\} \tag{A.6}$$

Summing up Eqs. A.2-A.6 results in $|k_\varepsilon(\xi_k, \xi_l) - k_\varepsilon(x_k, x_l)| \leq b_{a,\varepsilon,\kappa}$. □

A.5 Proof of Lemma 4.4

Lemma 4.4. *For any arbitrary choice of pairs* $(x_k, x_l) \in \Omega_k \times \Omega_l$ *and* $(\xi_k, \xi_l) \in \Omega_k \times \Omega_l$, $|k_\varepsilon(\xi_k, \xi_l) - k_\varepsilon(x_k, x_l)| \leq h_{\varepsilon,\kappa}$.

Proof. Due to Eq. 3.3

$$\begin{aligned}
|k_\varepsilon(\xi_k,\xi_l) - k_\varepsilon(\xi_k,x_l)| &= \left| \int g_\varepsilon(\xi_k - r) g_\varepsilon(\xi_l - r) q(r) dr \right. \\
&- \left. \int g_\varepsilon(\xi_k - r) g_\varepsilon(x_l - r) q(r) dr \right| \\
&= \left| \int [g_\varepsilon(\xi_l - r) - g_\varepsilon(x_l - r)] g_\varepsilon(\xi_k - r) q(r) dr \right| \\
&\leq \int |g_\varepsilon(\xi_l - r) - g_\varepsilon(x_l - r)| \, g_\varepsilon(\xi_k - r) q(r) dr \\
&\leq (\pi\varepsilon)^{-m/2} \sqrt{\frac{2}{e\varepsilon}} \|\xi_l - x_l\| \int g_\varepsilon(\xi_k - r) q(r) dr \\
&\leq \pi^{-m/2} \varepsilon^{-(m+1)/2} \kappa
\end{aligned}$$

Since $|k_\varepsilon(\xi_k,\xi_l) - k_\varepsilon(x_k,x_l)| \leq |k_\varepsilon(\xi_k,\xi_l) - k_\varepsilon(\xi_k,x_l)| + |k_\varepsilon(\xi_k,x_l) - k_\varepsilon(x_k,x_l)|$, we get $|m_{kl} - k_\varepsilon(x_k,x_l)| \leq h_{\varepsilon,\kappa}$. □

Acknowledgment

This research was partially supported by the Israel Science Foundation (Grant No. 1041/10), by the Israeli Ministry of Science & Technology (Grants No. 3-9096, 3-10898), by US - Israel Binational Science Foundation (BSF 2012282) and by a Fellowship from Jyväskylä University.

[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. *Submitted*, 2012.

[3] A. Bermanis, G. Wolf, and A. Averbuch. Measure-based diffusion kernel methods. In *SampTA 2013: 10th international conference on Sampling Theory and Applications*, Bremen, Germany, 2013.

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

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

[6] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[7] N. Djurdjevac, M. Sarich, and C. Schütte. Estimating the eigenvalue error of markov state models. *Multiscale Modeling & Simulation*, 10(1):61–81, 2012.

[8] 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.

[9] H. Guillard, A. Janka, and P. Vaněk. Analysis of an Algebraic Petrov–Galerkin Smoothed Aggregation Multigrid Method. *Appl. Numer. Math.*, 58(12):1861–1874, December 2008.

[10] M Horn. Optimal algorithms for global optimization in case of unknown Lipschitz constant. *Journal of Complexity*, 22(1):50 – 70, 2006. Special Issue Algorithms and Complexity for Continuous Problems Special Issue.

[11] G. Horton and S. T. Leutenegger. A multi-level solution algorithm for steady-state markov chains. In *in Proceedings of the ACM SIGMETRICS 1994 Conference on Measurement and Modeling of Computer Systems*, pages 191–200, 1994.

[12] W. Huisinga. *Metastability of Markovian systems*. PhD thesis, Freie Universität Berlin, 2001.

[13] M. Killian. *Algebraic Multigrid for Markov Chains and Tensor Decomposition*. PhD thesis, University of Waterloo, 2012.

[14] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

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

[16] J. Lin. Mapreduce is good enough? if all you have is a hammer, throw away everything that's not a nail! *CoRR*, abs/1209.2191, 2012.

[17] B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 955–962. MIT Press, Cambridge, MA, 2006.

[18] B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

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

[20] Marco Sarich, Frank Noé, and Christof Schütte. On the approximation quality of markov state models. *Multiscale Modeling & Simulation*, 8(4):1154–1177, 2010.

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

[22] E. Treister and I. Yavneh. On-the-fly adaptive smoothed aggregation multigrid for markov chains. *SIAM Journal on Scientific Computing*, 33(5):2927–2949, 2011.

[23] U.Fayyad. Big data analytics: Applications and opportunities in on-line predictive modeling, 2012. http://big-data-mining.org/keynotes/#fayyad.

[24] P. Vaněk, Y. Van Ek, A. Janka, , and H. Guillard. Convergence of algebraic multigrid based on smoothed aggregation ii: Extension to a petrov-galerkin method. *Computing*, 56:179–196, 1998.

[25] T. White. *Hadoop: The Definitive Guide: The Definitive Guide*. O'Reilly Media, 2009.

[26] G. Yang, X. Xu, and J. Zhang. Manifold alignment via local tangent space alignment. *International Conference on Computer Science and Software Engineering*, December 2008.

[27] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *Technical Report CSE-02-019, Department of Computer Science and Engineering, Pennsylvania State University*, 2002.