Journal of Nonlinear Science (2021) 31:3
https://doi.org/10.1007/s00332-020-09668-z

Journal of Nonlinear Science

Dimensionality Reduction of Complex Metastable Systems via Kernel Embeddings of Transition Manifolds

Andreas Bittracher[1] · Stefan Klus[1] · Boumediene Hamzi[2,3] · Péter Koltai[1] · Christof Schütte[1,4]

Received: 31 May 2019 / Accepted: 12 November 2020 / Published online: 18 December 2020

Abstract

We present a novel kernel-based machine learning algorithm for identifying the low-dimensional geometry of the effective dynamics of high-dimensional multiscale stochastic systems. Recently, the authors developed a mathematical framework for the computation of optimal reaction coordinates of such systems that is based on learning a parameterization of a low-dimensional transition manifold in a certain function space. In this article, we enhance this approach by embedding and learning this transition manifold in a reproducing kernel Hilbert space, exploiting the favorable properties of kernel embeddings. Under mild assumptions on the kernel, the manifold structure is shown to be preserved under the embedding, and distortion bounds can be derived. This leads to a more robust and more efficient algorithm compared to the previous parameterization approaches.

1 Introduction

Many of the dynamical processes investigated in the sciences today are characterized by the existence of phenomena on multiple, interconnected time scales that determine the long-term behavior of the process. Examples include the inherently multiscale

Communicated by Oliver Junge.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00332-020-09668-z.

✉ Andreas Bittracher
bittracher@mi.fu-berlin.de

1 Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany

2 Department of Mathematics, Imperial College London, London, UK

3 AlFaisal University, Riyadh, Kingdom of Saudi Arabia

4 Zuse Institute Berlin, Berlin, Germany

dynamics of atmospheric vortex- and current formation which needs to be considered for effective weather prediction (Klein 2010; Majda and Klein 2003), or the vast difference in time scales on which bounded atomic interactions, side-chain interactions, and the resulting formation of structural motifs occur in biomolecules (Freddolino et al. 2010; Camacho and Thirumalai 1993; Bowman et al. 2011). An effective approach to analyze these systems is often the identification of a low-dimensional observable of the system that captures the interesting behavior on the longest time scale. However, the computerized identification of such observables from simulation data poses a significant computational challenge, especially for high-dimensional systems.

Recently, the authors have developed a novel mathematical framework for identifying such essential observables for the slowest time scale of a system (Bittracher et al. 2017). The method—called the *transition manifold approach*—was primarily motivated by molecular dynamics, where the dynamics is typically described by a thermostated Hamiltonian system or diffusive motion in molecular dynamics landscapes. In these systems, local minima of the potential energy landscape induce *metastable behavior*, which is the phenomenon that on long time scales, the dynamics is characterized by rare transitions between certain sets that happen roughly along interconnecting *transition pathways* (Noé et al. 2009; Schütte et al. 2013; E and Vanden-Eijnden 2006). The sought-after essential observables should thus resolve these transition events, and are called *reaction coordinates* in this context (Socci et al. 1996; Best and Hummer 2005), a notion that we will adopt here. Despite of its origins, the transition manifold approach is also applicable to other classes of reducible systems.

At the heart of this approach is the insight that good reaction coordinates can be found by parameterizing a certain *transition manifold* $\mathbb{M}$ in the function space L^1. For metastable systems, this manifold has strong connections to the aforementioned transition pathway (Bittracher et al. 2018), but the two concepts are not equivalent. Its defining property is that for times τ that fall between the fastest and slowest time scales, the *transition density functions* with relaxation time τ concentrate around $\mathbb{M}$. Hence, $\mathbb{M}$ can be seen as the "backbone" of the slowly equilibrating parts of the dynamics.

The original algorithmic strategy to compute an RC by parameterizing $\mathbb{M}$, proposed in Bittracher et al. (2017) can be summarized as follows:

1. Randomly choose starting points in the dynamically relevant regions of the state space.
2. Approximate the transition densities associated with each starting point by Monte Carlo simulation.
3. Embed the transition densities into a Euclidean space by a suitable embedding function.
4. Uncover the manifold structure in the set of embedded transition densities with the help of some manifold learning algorithm.

The result is a reaction coordinate evaluated in the starting points. For an appropriately chosen embedding function, this reaction coordinate has been shown to be as expressive as the dominant eigenfunctions of the transfer operator associated with the system (Bittracher et al. 2017), which can be considered an "optimal" reaction coordinate (Froyland et al. 2016; Bowman et al. 2014; McGibbon et al. 2017). One decisive advantage of the transition manifold reaction coordinates over the eigenfunctions,

however, is the ability to compute the reaction coordinate *locally* (by choosing the starting points), whereas with conventional methods, the inherently global computation of transfer operator eigenfunctions quickly becomes infeasible due to the curse of dimensionality [although kernel-based methods alleviate this problem to some extent (Schwantes and Pande 2015; Klus et al. 2018)]. Moreover, the number of dominant eigenfunctions can be significantly larger than the "natural" dimension of the reaction coordinate, which the transition manifold method successfully discovers (Bittracher et al. 2017).

Despite the success of the original framework in defining and computing dynamically verifiable reaction coordinates, the original algorithm had several shortcomings related to the choice of the embedding function. First, in order to ensure the preservation of the manifold's topology under the embedding, the dimension of $\mathbb{M}$ had to be known in advance. Second, the particular way of choosing the embedding functions allowed no control over the distortion of $\mathbb{M}$ under the embedding, which may render the parameterization problem numerically ill-conditioned.

The goal of this article is to overcome both of these problems by *kernelizing* the transition manifold embedding. That is, we present a method to implicitly embed the transition manifold into a *reproducing kernel Hilbert space* (RKHS) with a proper kernel, instead of embedding it into a Euclidean space. The RKHS is—depending on the kernel—a high- or even infinite-dimensional function space with the crucial property that inner products between functions embedded into it can be computed by cheap kernel evaluations, without ever explicitly having to compute the embedding (Steinwart and Christmann 2008; Schölkopf and Smola 2001), something that is known as the *kernel trick*. In our case, this means that the pairwise distance between embedded transition densities—a key component in the manifold learning part of the algorithm—can be computed efficiently by kernel evaluations at samples of the densities.

Due to their popularity, the metric properties of the kernel embedding are well-studied (Smola et al. 2007; Fukumizu et al. 2007; Sriperumbudur et al. 2010; Gretton et al. 2012; Muandet et al. 2017). In particular, for characteristic kernels, the RKHS is "large" in an appropriate sense, and geometrical information is well-preserved under the embedding. For our application, this will mean that distances between points on the transition manifold $\mathbb{M}$ are approximately preserved, and thus the distortion of $\mathbb{M}$ under the embedding can be bounded. Also, such a "large" RKHS can embed transition manifolds of arbitrary finite dimension, hence a priori knowledge of the dimension of $\mathbb{M}$ is no longer required.

In a more general machine learning context, the kernel trick is often used to derive nonlinear versions of originally linear algorithms, by interpreting the RKHS-embedding of a data set as a high-dimensional, nonlinear transformation, and (implicitly) applying the linear algorithm to the transformed data. This approach has been successfully applied to methods such as *principal component analysis* (PCA) (Schölkopf et al. 1998), *canonical correlation analysis* (CCA) (Melzer et al. 2001), and *time-lagged independent component analysis* (TICA) (Schwantes and Pande 2015), to name but a few. Transferred to our application this means that, if the transformation induced by the kernel embedding is able to approximately linearize the transition manifold, there is hope that efficient *linear* manifold learning methods can be used to parameterize the embedded transition manifold.

The main contributions and the structure of this article are as follows: In Sect. 2, we will revisit the definition of transition manifolds and discuss conditions under which systems possess such manifolds. Also, the old algorithm based on Euclidean embeddings is revisited here. Section 3 constitutes the main part of this article. In Sect. 3.1, reproducing kernel Hilbert spaces and the kernel trick are introduced. Also, our new algorithm is derived by reformulating the old algorithm using the kernel trick. In Sect. 3.2, we derive bounds for the maximum possible distortion of the transition manifold under the kernel embedding. These bounds provide insight into the condition and well-posedness of our new method. In Sect. 4, the performance of the algorithm is evaluated by three examples. In Sect. 4.1, the kernel reaction coordinate is computed for a two-dimensional standard benchmark system (the Müller–Brown potential), and compared to an "optimal" reaction coordinate. In Sect. 4.2, the distortions of the transition manifold under the Euclidean and the kernel embedding are compared using a two-dimensional toy system. In Sect. 4.3, the method is applied to a 66-dimensional peptide system, revealing the transition pathways between metastable conformations. Finally, a conclusion and comments on future work are given in Sect. 5.

2 Reaction Coordinates Based on Transition Manifolds

In what follows, let $\{X_t\}_{t\geq 0}$ (abbreviated as X_t) be a reversible, ergodic, stochastic process on a compact connected state space $\mathbb{X} \subset \mathbb{R}^n$ with positive, finite Lebesgue measure. Let there furthermore exist a unique invariant density $\rho \in L^1(\mathbb{X})$, $0 < \rho < \infty$ of X_t, i.e., if $X_0 \sim \rho$, then $X_t \sim \rho$ for all $t \geq 0$. $L^1(\mathbb{X})$ here denotes the space of absolutely integrable functions over $\mathbb{X}$ with respect to the Lebesgue measure. For instance, a process generated by a stochastic differential equation (SDE) with sufficiently smooth parameters, uniformly non-degenerate noise coefficient, and a domain $\mathbb{X}$ with sufficiently smooth boundary fulfills these requirements. Typical classes of these SDEs are (overdamped) Langevin equations, where ρ takes the form of the Boltzmann–Gibbs distribution. See, e.g., (Mattingly and Stuart 2002), for related statements.

For fixed $x \in \mathbb{X}$ and $t > 0$, let $p_x^t : \mathbb{X} \to \mathbb{R}_{>0}$ denote the transition density function of the system, i.e., p_x^t describes the probability density at time t, after having started in point x at time 0. Under sufficient smoothness requirements on the system, p_x^t is indeed a function, and we will often consider p_x^t as a point in $L^1(\mathbb{X})$, and later also other related function spaces. For the sake of clarity, we will from now on omit the argument of L^p when referring to functions over $\mathbb{X}$.

2.1 Reducibility of Dynamical Systems

We assume the state space dimension n to be large. The main objective of this work is the identification of good low-dimensional *reaction coordinates* (RCs) or *order parameters* of the system. An r-dimensional RC is a smooth map $\xi : \mathbb{X} \to \mathbb{Y}$ from the full state space $\mathbb{X}$ to a lower-dimensional space $\mathbb{Y} \subset \mathbb{R}^r$, $r \ll n$. Loosely speaking, we call such an RC *good* if on long enough time scales the projected process

$\xi(X_t)$ is approximately Markovian and the dominant spectral properties of the operator describing its density evolution of $\xi(X_t)$ resemble those of X_t. This ensures that important long-time statistical properties such as equilibration times are preserved under projection onto the RC.

We will now introduce a conceptual framework for finding such RCs. This so-called *transition manifold framework*, introduced by some of the authors in Bittracher et al. (2017), ties the existence of good RCs to certain geometrical properties of the following family of transition densities:

Definition 2.1 Let $\tau > 0$ be fixed. The set of functions

$$\widetilde{\mathbb{M}} := \{p_x^\tau \mid x \in \mathbb{X}\} \subset L^1,$$

is called the *fuzzy transition manifold* of the system.

The name *fuzzy transition manifold* is motivated by the following observation: If and only if for some lag time τ the transition density functions p_x^τ do not depend on the full coordinate x, but only depend smoothly on some r-dimensional reaction coordinate $\xi(x)$, i.e.,

$$p_x^\tau = \widetilde{p}_{\xi(x)}^\tau$$

for all $x \in \mathbb{X}$ and some smooth injective function $\widetilde{p}_{(\cdot)}^\tau : \mathbb{Y} \to L^1(\mathbb{X})$, then $\widetilde{\mathbb{M}}$ will form an r-dimensionally parameterizable set in L^1. In a slight misuse of denotation, we will call such a set an *r-dimensional manifold* for short, and whenever we talk of a manifold, we mean this special type of manifold (i.e., extrinsically defined by a parameterization), and not a general topological manifold. Likewise, if and only if p_x^τ "almost depends" only on $\xi(x)$, i.e.,

$$p_x^\tau \approx \widetilde{p}_{\xi(x)}^\tau, \tag{1}$$

then the $\widetilde{\mathbb{M}}$ will "almost form" an r-dimensional manifold. More precisely, $\widetilde{\mathbb{M}}$ will then cluster around some actual r-dimensional manifold $\mathbb{M} \subset L^1$, i.e., $\widetilde{\mathbb{M}}$ will be close to $\mathbb{M}$ in some appropriate metric. This leads to the following definition:

Definition 2.2 The process X_t is called (ε, r)-*reducible* if there exists an r-dimensional manifold of functions $\mathbb{M} \subset \widetilde{\mathbb{M}}$ such that for a fixed lag time τ, it holds that

$$\min_{f \in \mathbb{M}} \|f - p_x^\tau\|_{L^2_{1/\rho}} \leq \varepsilon \quad \text{for all } x \in \mathbb{X}. \tag{2}$$

Any such $\mathbb{M}$ is called a *transition manifold* (TM) of the system.

The algorithmic idea now is based on the inverse problem: suppose (2) holds, and we know (or are able to compute) a homeomorphic parameterization $\mathcal{E}$ of $\mathbb{M}$, i.e., $\mathcal{E} : \mathbb{M} \to \mathbb{Y} \subset \mathbb{R}^r$. Then, as we will describe in detail in Sect. 2.3, an RC ξ fulfilling (1) can be constructed from $\mathcal{E}$, and such a ξ indeed preserves the slowest time scales of the system.

The remaining question is under which conditions a process X_t is (ε, r)-reducible. Unfortunately, a full characterization is still missing. The general intuition is that in time scale separated systems, the r-dimensional structure in $\widetilde{\mathbb{M}}$ emerges with progressive equilibration of the fast time scales. Let δ_x denote the Dirac distribution centered in x. For $\tau \to 0$, we have $\widetilde{\mathbb{M}} \to \{\delta_x \mid x \in \mathbb{X}\}$ (in the sense of distributions), which is one-to-one to $\mathbb{X}$, an n-dimensional space. On the other hand, for $\tau \to \infty$, we have $\widetilde{\mathbb{M}} \to \{\rho\}$, as every p_x^τ converges to the invariant density ρ. If the system now consists of r slowly-equilibrating "components", and $n - r$ quickly-equilibrating "components", separated by a significant time scale gap, then $\widetilde{\mathbb{M}}$ must go through a phase of being almost r-dimensional on its path from being n- to being 0-dimensional.

It is hard to pinpoint what exactly "component" here means, and the meaning varies with the class of the system under consideration. On the one hand, the explicitly time scale separated systems, following the SDE

$$\begin{aligned} \mathrm{d}X_t^{(1)} &= f(X_t^{(1)}, X_t^{(2)})\mathrm{d}t + \mathrm{d}W_t^{(1)}, \\ \mathrm{d}X_t^{(2)} &= \frac{1}{\kappa} g(X_t^{(1)}, X_t^{(2)})\mathrm{d}t + \frac{1}{\sqrt{\kappa}}\mathrm{d}W_t^{(2)} \end{aligned}$$

with small parameter $0 < \kappa \ll 1$, possess the slow variable $x^{(1)} \in \mathbb{R}^r$, and the fast variable $x^{(2)} \in \mathbb{R}^{n-r}$. Here, we conjecture that for large enough τ, p_x^τ essentially depends only on $\xi(x) = x^{(1)}$, hence an r-dimensional TM exists. While this statement can be easily understood for concrete systems, a general formulation and proof is still lacking.

For metastable systems, on the other hand, one common misconception is that the number of slow components corresponds to the number of dominant eigenvalues of the Perron–Frobenius operator (see Sect. 2.2 for its definition), and that the RC corresponds to the dominant eigenfunctions. However, it has been demonstrated in Bittracher et al. (2017), Appendix B, that the number of dominant eigenvalues is only an upper bound for r. Moreover, the RC is suspected to be more closely related to a parameterization of the network of transition pathways that connects the metastable sets, than to the eigenfunctions. While again we do not have rigorous proof of that conjecture, or of the existence of a TM in general metastable systems, the following two-dimensional example supports the claim that both are reasonable assumptions: stop

Example 2.3 Consider the process X_t to be described by overdamped Langevin dynamics

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2/\beta}\mathrm{d}W_t, \tag{3}$$

with the energy potential V, the inverse temperature β, and Brownian motion W_t. The potential depicted in Fig. 1 (left) possesses two metastable states, located around the two local energy minima, that are connected by a one-dimensional transition path.

If the temperature is sufficiently low, and the lag time τ is high enough, the probability to find the system at time τ outside of the metastable sets is miniscule, for each starting point. Hence, for each x, the transition density p_x^τ is essentially a convex com-

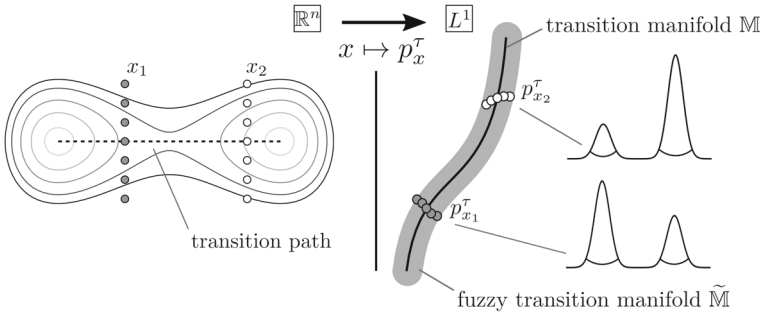

Fig. 1 Illustration of a metastable system and its transition manifold. Left: Two-dimensional energy potential with two metastable sets. Right: Relative positions of the transition densities to each other. The densities for each gray starting point resemble each other, hence they concentrate around one point in L^1. The same holds for the white starting points. Overall, the densities p_x^τ vary only substantially with the progress of x along the transition pathway. Hence, the set $\widetilde{\mathbb{M}}$ concentrates around a one-dimensional manifold in L^1

bination of the two quasi-stationary densities[1] of the metastable sets. Moreover, the convex factor here only depends on the horizontal coordinate of x, i.e., $\xi(x) = x^{(1)}$, as the probability of whether a trajectory will be "caught" by the left or right well depends almost exclusively on the progress of x along the transition path, which can be described by the horizontal coordinate. Hence, we have

$$p_x^\tau \approx \widetilde{p}_{\xi(x)}^\tau$$

for some functions $\widetilde{p}_{(\cdot)}^\tau$, and thus $\widetilde{\mathbb{M}}$ concentrates around a one-dimensional manifold $\mathbb{M}$ in L^1.

Due to the aforementioned difficulties in connecting the existence of a TM to more conventional conditions for reducibility, we will in the following always directly assume that the process X_t is (ε, r)-reducible with small ε and $r \ll n$.

Two technical remarks regarding Definition 2.2 are in order:

1. Note that in the above definition, the $L^2_{1/\rho}$ norm is used to measure distances, where $L^2_{1/\rho}$ is the space of (equivalence classes of) functions that are square-integrable with respect to the measure induced by the function $\frac{1}{\rho}$, and thus for $f \in L^2_{1/\rho}$,

$$\|f\|_{L^2_{1/\rho}} = \left(\int_{\mathbb{X}} f(x)^2 \frac{1}{\rho(x)} \mathrm{d}x \right)^{1/2}.$$

   Closeness with respect to the $L^2_{1/\rho}$-norm instead of the L^1-norm is indeed a strict requirement here, as measuring the quality of a given RC will require a Hilbert space, see Sect. 2.2. Note that under appropriate assumptions on the system, it holds that $p_x^t \in L^2_{1/\rho}$ for all $x \in \mathbb{X}$. This will be shown in Lemma 3.13 and implies $\widetilde{\mathbb{M}} \subset L^2_{1/\rho}$, which together with the requirement $\mathbb{M} \subset \widetilde{\mathbb{M}}$ makes (2) well-defined.

[1] The quasi-stationary density of a set A is the equilibrium density of the system X_t conditioned on remaining inside A for all future times (Gesùa et al. 2016).

2. The original definition of (ε, r)-reducibility (see Bittracher et al. 2017, Definition 4.4), is marginally different from the definition above: Instead of $\mathbb{M} \subset L^2_{1/\rho}$, we here require $\mathbb{M} \subset \widetilde{\mathbb{M}} \subset L^2_{1/\rho}$. The introduction of this slightly stronger technical requirement allows us to later control a certain embedding error, see Proposition 2.7. Note that the proofs in Bittracher et al. (2017) regarding the optimality of the final reaction coordinate are not affected by this change.

2.2 A Measure for the Quality of Reaction Coordinates

We will now present a measure for evaluating the quality of reaction coordinates that is based on transfer operators, first derived in Bittracher et al. (2017). The Perron–Frobenius operator $\mathcal{P}^t : L^1 \to L^1$ associated with the process X_t is defined by

$$\left(\mathcal{P}^t u\right)(y) = \int_{\mathbb{X}} u(x)\, p_x^t(y)\, \mathrm{d}x.$$

This operator can be seen as the push-forward of arbitrary starting densities, i.e., if $X_0 \sim u$, then $X_t \sim \mathcal{P}^t u$.

As $L^2_{1/\rho} \subset L^1$ (see Bittracher et al. 2017, Remark 4.6) we can consider $\mathcal{P}^t$ as an operator on the inner product space $L^2_{1/\rho}$, where it has particularly advantageous properties (see Baxter and Rosenthal 1995; Schervish and Carlin 1992; Klus et al. 2018). Here, it is self-adjoint due to the reversibility of X_t. Moreover, under relatively mild conditions, it does not exhibit any essential spectrum (Schütte et al. 2013). Hence, its eigenfunctions form an orthonormal basis of $L^2_{1/\rho}$ and the associated eigenvalues are real. Now, the significance of the dominant eigenpairs for the system's time scales is well-known (Schütte et al. 2013). This is the primary reason for the choice of the $L^2_{1/\rho}$-norm in Definition 2.2.

Let θ_i^t be the eigenvalues of $\mathcal{P}^t$, sorted by decreasing absolute value, and ψ_i the corresponding eigenfunctions, where $i = 0, 1, \ldots$. It holds that $\theta_0 = 1$ is independent of t, isolated and the sole eigenvalue with absolute value 1. Furthermore, $\psi_0 = \rho$. The subsequent eigenvalues decrease monotonously to zero both for increasing index and time. That is,

$$\lim_{i\to\infty} |\theta_i^t| = 0 \quad \text{and} \quad \lim_{t\to\infty} |\theta_i^t| = 0.$$

The associated eigenfunctions $\psi_1, \psi_2, \ldots$ can be interpreted as sub-processes of decreasing longevity in the following sense: Let $u \in L^2_{1/\rho}$, with $u = \sum_{i=0}^{\infty} \alpha_i \psi_i$, $\alpha_i \in \mathbb{R}$, then

$$\mathcal{P}^t u = \sum_{i=0}^{\infty} \theta_i^t \alpha_i \psi_i \approx \sum_{i=0}^{d} \theta_i^t \alpha_i \psi_i$$

since for the lag time $\tau > 0$ as defined above, there exists an index $d \in \mathbb{N}$ such that $|\theta_i^t| \approx 0$ for all $t \geq \tau$ and all $i > d$. Hence, the major part of the information about the long-term density propagation of X_t is encoded in the d dominant eigenpairs.

The operator $\mathcal{P}^t$ describes the evolution of densities of the full process X_t. In order to monitor the dependence of densities on the reduced coordinate ξ only, we first introduce the projection operator $\Pi_\xi : L^1(\mathbb{X}) \to L^1(\mathbb{X})$,

$$\big(\Pi_\xi(u)\big)(y) = \mathbb{E}_\rho\big[u(x) \;\big|\; x \in \xi^{-1}\big(\xi(y)\big)\big], \tag{4}$$

i.e., we take the expectation value of u with respect to ρ on the $\xi(y)$-levelset. Intuitively, Π_ξ averages a function $u \in L^1(\mathbb{X})$ over the individual level sets of ξ, hence $\Pi_\xi u$ is constant on each level set of ξ. Π_ξ is equivalent to the Zwanzig projection operator from statistical physics (Zwanzig 2001, Nov 1961), although the latter is typically constructed as a map into $L^1(\mathbb{Y})$. We however require Π_ξ to map into $L^1(\mathbb{X})$ to be able to directly compare its input and output functions. For a detailed investigation of Π_ξ, see Zhang et al. (2016) and Bittracher et al. (2017).

Using Π_ξ, the *effective transfer operator* $\mathcal{P}_\xi^t : L^1(\mathbb{X}) \to L^1(\mathbb{X})$ associated with ξ is then given by

$$\mathcal{P}_\xi^t u = \Pi_\xi\big(\mathcal{P}^t(\Pi_\xi u)\big),$$

see Bittracher et al. (2017). We now want to preserve the statistics of the dominant long-term dynamics of X_t under the projection onto ξ, i.e.,

$$\mathcal{P}^t u \approx \mathcal{P}_\xi^t u, \tag{5}$$

for $t \geq \tau$, where τ is some lag time that is long enough for the fast processes, associated with the non-dominant eigenpairs, to have equilibrated. A sufficient condition for (5) is

$$\Pi_\xi \psi_i \approx \psi_i, \quad i = 0, \ldots, d,$$

that is, the dominant eigenfunctions ψ_i must be almost constant along the level sets of ξ. This motivates the following definition of a good reaction coordinate:

Definition 2.4 Let (ψ_i, θ_i^t) be the eigenpairs of the Perron–Frobenius operator. Let $\tau > 0$ and $d \in \mathbb{N}$ such that $\theta_i^t \approx 0$ for all $i > d$ and $t \geq \tau$. We call a function $\xi : \mathbb{X} \to \mathbb{R}^r$ a *good reaction coordinate* if for all $i = 0, \ldots, d$ there exist functions $\tilde{\psi}_i : \mathbb{R}^r \to \mathbb{R}$ such that

$$\big\| \psi_i - \tilde{\psi}_i \circ \xi \big\|_\infty \approx 0. \tag{6}$$

If condition (6) is fulfilled, we say that ξ (approximately) *parameterizes* the dominant eigenfunctions.

For a formal evaluation of the condition (6), see Bittracher et al. (2017), Corollary 3.6).

2.3 Optimal Reaction Coordinates

We now justify why reaction coordinates that are based on parameterizations of the transition manifold $\mathbb{M}$ indeed fulfill condition (6). Let $Q\colon L^2_{1/\rho} \to L^2_{1/\rho}$ be the nearest-point projection onto $\mathbb{M}$, i.e.,

$$Q(f) = \arg\min_{g\in\mathbb{M}} \|f - g\|_{L^2_{1/\rho}}.$$

Assume further that some parameterization $\gamma\colon \mathbb{M} \to \mathbb{R}^r$ of $\mathbb{M}$ is known, i.e., γ is one-to-one on $\mathbb{M}$ and its image in $\mathbb{R}^r$. Then the reaction coordinate $\xi\colon \mathbb{R}^n \to \mathbb{R}^k$ defined by

$$\xi(x) := \big(\gamma \circ Q\big)(p^\tau_x) \tag{7}$$

is good in the sense of Definition 2.4 due to the following theorem:

Theorem 2.5 (Bittracher et al. 2017, Corollary 3.8). *Let the system be (ε, r)-reducible and ξ defined as in (7). Then for all $i = 0, \dots, d$, there exist functions $\tilde{\psi}_i\colon \mathbb{R}^r \to \mathbb{R}$ such that*

$$\big\|\psi_i - \tilde{\psi}_i \circ \xi\big\|_\infty \leq \frac{\varepsilon}{|\theta^\tau_i|}. \tag{8}$$

Let us add two remarks:

1. The choice of the $L^2_{1/\rho}$-norm in Definition 2.2 is crucial for Theorem 2.5 to hold.
2. Metastable systems typically exhibit a time scale gap after the d dominant eigenvalues, i.e.,

$$\frac{|\theta^t_d - \theta^t_{d+1}|}{|\theta^t_{d+1} - \theta^t_{d+2}|} \gg 1 \quad \text{for suitably large } t > 0.$$

In this case, τ can be chosen such that $|\theta^\tau_{d+1}|$ is close to zero and $|\theta^\tau_i|$, $i = 0, \dots, d$, is still relatively large. Consequently, the denominator in (8) is not too small, and thus the RC (7) is indeed good according to Definition 2.4.

The main task for the rest of the paper is now the numerical computation of an (approximate) parameterization γ of $\mathbb{M}$.

2.4 Whitney Embedding of the Transition Manifold

One approach to find a parameterization of $\mathbb{M}$, proposed by the authors in Bittracher et al. (2017), is to first embed $\mathbb{M}$ into a more accessible Euclidean space and to parameterize the embedded manifold. In order to later compare it with our new method, we will briefly describe this approach here.

To construct an embedding $\mathcal{E}$ that preserves the topological structure of $\mathbb{M}$, without prior knowledge about $\mathbb{M}$, a variant of the Whitney embedding theorem can be used. It extends the classic Whitney theorem to arbitrary Banach spaces and was proven by Hunt and Kaloshin in Hunt and Kaloshin (1999).

Theorem 2.6 [Whitney embedding theorem in Banach spaces, (Hunt and Kaloshin 1999)]. *Let $\mathbb{V}$ be a Banach space and let $\mathbb{K} \subset \mathbb{V}$ be a manifold of dimension r. Let $k > 2r$ and let $\alpha_0 = \frac{k-2d}{k(d+1)}$. Then, for all $\alpha \in (0, \alpha_0)$, for almost every (in the sense of prevalence) bounded linear map $\mathcal{F}: \mathbb{V} \to \mathbb{R}^k$ there exists a $C > 0$ such that for all $x, y \in \mathbb{K}$,*

$$C \, \|\mathcal{F}(x) - \mathcal{F}(y)\|_2^{\alpha} \geq \|x - y\|_{\mathbb{V}} \,,$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^k$. In particular, almost every $\mathcal{F}$ is one-to-one on $\mathbb{K}$ and its image, and $\mathcal{F}^{-1}\big|_{\mathcal{F}(\mathbb{K})}$ is Hölder continuous with exponent α.

In particular, almost every such map $\mathcal{F}$ is a homeomorphism between $\mathbb{K}$ and its image in $\mathbb{R}^k$, which in short is called an *embedding* of $\mathbb{K}$ (see e.g. Munkres 2000, §18). This means that the image $\mathcal{F}(\mathbb{M})$ will again be an r-dimensional manifold in $\mathbb{R}^k$, provided that $k > 2r$. We will apply this result to the transition manifold, i.e., $\mathbb{V} = L^2_{1/\rho}$ and $\mathbb{K} = \mathbb{M}$, and for simplicity restrict ourselves to the lowest embedding dimension, i.e., $k = 2r + 1$. Any "randomly selected" continuous map $\mathcal{F} : L^2_{1/\rho} \to \mathbb{R}^{2r+1}$ then is an embedding of $\mathbb{M}$.

Unfortunately, there is no practical way to randomly draw from the space of continuous maps on $L^2_{1/\rho}$ directly. Instead of arbitrary continuous maps, we therefore restrict our considerations to maps $\mathcal{F}: L^2_{1/\rho} \to \mathbb{R}^{2r+1}$ of the form

$$\mathcal{F}(f) := \int_{\mathbb{X}} \eta(x') f(x') \mathrm{d}x', \tag{9}$$

where

$$\eta(x) := Ax, \quad A \in \mathbb{R}^{(2r+1)\times d}, \; A \sim \sigma,$$

where σ is some distribution on the (finite-dimensional) space of $(2r+1) \times d$-matrices (e.g., Gaussian matrices). The linear map $\eta: \mathbb{X} \to \mathbb{R}^{2r+1}$, called *feature map*, is bounded due to the boundedness of $\mathbb{X}$. Maps of the form (9) are therefore continuous on L^1, and thus in particular on the subspace $L^2_{1/\rho}$.

By drawing from the distribution σ of the matrices A, we can effectively sample maps of form (9). There is however no formal guarantee that maps of form (9) fall into the prevalent set of maps predicted by Theorem 2.6, and for general manifolds $\mathbb{K} \subset L^2_{1/\rho}$, this is indeed not the case.[2] However, there is empirical evidence that transition manifolds in real-world systems, specifically molecular dynamical systems, are "sufficiently regular" for (9) to preserve their features well (Bittracher et al. 2018). Still, this necessary restriction to a finite-dimensional class of embedding functions represents a significant deficit of the framework described here. We will see later how this deficit can be resolved by instead using embeddings based on kernel functions.

[2] For example, if $\mathbb{K}$ is an r-dimensional linear subspace of $L^2_{1/\rho}$ spanned by basis functions with identical expectation values, then no $\mathcal{F}$ of form (9) will embed $\mathbb{K}$.

Still, for the moment, we assume that a randomly drawn function of form (9) with linear η almost surely is an embedding of $\mathbb{M}$. The *dynamical embedding* of a point $x \in \mathbb{X}$ is then defined by

$$\mathcal{E}(x) := \mathcal{F}(p_x^t) = \int \eta(x') p_x^t(x') \, \mathrm{d}x'. \tag{10}$$

This is the Euclidean representation of the density p_x^t, and the set $\{\mathcal{E}(x) \mid x \in \mathbb{X}\} \subset \mathbb{R}^{2r+1}$ is the Euclidean representation of the fuzzy transition manifold. It again clusters around an r-dimensional manifold in $\mathbb{R}^{2r+1}$, namely the image $\mathcal{F}(\mathbb{M})$ of the transition manifold under $\mathcal{F}$:

Proposition 2.7 *Let the process X_t be (ε, r)-reducible with transition manifold $\mathbb{M}$, and $\mathcal{F}\colon L^2_{1/\rho} \to \mathbb{R}^{2r+1}$ and $\mathcal{E}\colon \mathbb{R}^n \to \mathbb{R}^{2r+1}$ defined as in (9) and (10). Then*

$$\inf_{v \in \mathcal{F}(\mathbb{M})} \|v - \mathcal{E}(x)\|_\infty \leq \|\eta\|_\infty \varepsilon \quad \textit{for all } x \in \mathbb{X}.$$

Proof Let $x \in \mathbb{X}$. By the (ε, r)-reducibility of X_t (Definition 2.2), and the fact that $\mathbb{M} \subset \widetilde{\mathbb{M}}$, i.e., $\mathbb{M}$ itself consists of transition densities, there exists an $x^* \in \mathbb{X}$ such that $p_{x^*}^t \in \mathbb{M}$ and $\|p_x^t - p_{x^*}^t\|_{L^2_{1/\rho}} \leq \varepsilon$. Thus we have

$$\begin{aligned}
\inf_{v \in \mathcal{F}(\mathbb{M})} \|v - \underbrace{\mathcal{E}(x)}_{=\mathcal{F}(p_x^t)}\|_\infty &\leq \|\mathcal{F}(p_{x^*}^t) - \mathcal{F}(p_x^t)\|_\infty \\
&= \left\| \int_{\mathbb{X}} \eta(x') \left(p_{x^*}^t(x') - p_x^t(x')\right) \mathrm{d}x' \right\|_\infty \\
&\leq \|\eta\|_\infty \underbrace{\|p_{x^*}^t - p_x^t\|_{L^1}}_{\leq \|p_{x^*}^t - p_x^t\|_{L^2_{1/\rho}}} \leq \|\eta\|_\infty \varepsilon,
\end{aligned}$$

where $\|\cdot\|_{L^1} \leq \|\cdot\|_{L^2_{1/\rho}}$ was derived in Bittracher et al. (2017), Remark 4.6. □

Remark 2.8 Together, Theorem 2.6 and Proposition 2.7 guarantee at least a minimal degree of well-posedness of the embedding problem: The embedded manifold $\mathcal{F}(\mathbb{M})$ has the same topological structure as $\mathbb{M}$, and $\mathcal{F}(\widetilde{\mathbb{M}})$ clusters closely around it (if $\|\eta\|_\infty$ is small). However, guarantees on the *condition number* of the problem cannot be made. The manifold $\mathbb{M}$ will in general be distorted by $\mathcal{F}$, to a degree that might pose problems for numerical manifold learning algorithms. This problem is illustrated in Fig. 2. Such a situation typically occurs if some of the components of the embedding $\mathcal{F}$ are strongly correlated.

Additionally, the Whitney embedding theorem cannot guarantee that the *fuzzy* transition manifold $\widetilde{\mathbb{M}}$ will be preserved under the embedding, as analytically $\widetilde{\mathbb{M}}$ is not a manifold. Thus, $\mathcal{F}$ is in general not injective on $\widetilde{\mathbb{M}}$.

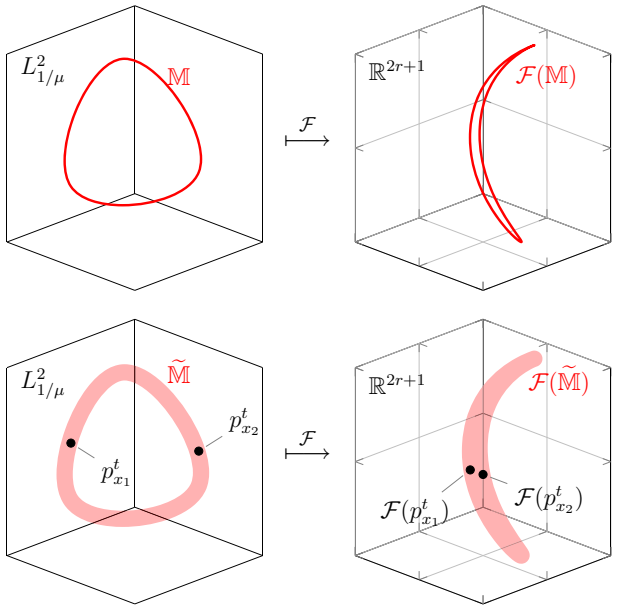

Fig. 2 Illustration of the consequences of bad choices for the embedding function. While the topology of the transition manifold $\mathbb{M}$ is preserved under the embedding, the relative distances between its points may be heavily distorted (top row). Intuitively, points that lie on distant parts of the manifold might be mapped closely together. As a consequence, a manifold learning algorithm based on distances between a finite number of samples of $\mathcal{F}(\widetilde{\mathbb{M}})$ would have difficulties learning the (in this case circular) topology of $\mathbb{M}$ (bottom row)

2.5 Data-Driven Algorithm for Parameterizing the Transition Manifold

Due to the implicit definition of $\mathbb{M}$, the embedded transition manifold $\mathcal{F}(\mathbb{M})$ is hard to analyze directly. However, as $\mathbb{M} \subset \widetilde{\mathbb{M}}$ and $\mathcal{F}(\widetilde{\mathbb{M}})$ concentrates $\big(\|\eta\|_\infty \varepsilon\big)$-closely around $\mathcal{F}(\mathbb{M})$, one can expect that any parameterization of the dominant directions of $\mathcal{F}(\widetilde{\mathbb{M}})$ is also a good parameterization of $\mathcal{F}(\mathbb{M})$. We now explain how $\mathcal{F}(\widetilde{\mathbb{M}})$ can be sampled numerically and how this sample can be parameterized.

Let $\mathbb{X}_N = \{x_1, \ldots, x_N\}$ be a finite sample of state space points, which covers the "dynamically relevant" part of state space, i.e., the regions of $\mathbb{X}$ of substantial measure ρ. The exact distribution of the sampling points is not important here. If $\mathbb{X}$ is bounded or periodic, $\mathbb{X}_N$ could be drawn from the uniform distribution or chosen to form a regular grid. In practice, it often consists of a sample of the system's equilibrium measure ρ.

The set $\mathcal{F}\big(\{p^t_x \mid x \in \mathbb{X}_N\}\big)$ will serve as our sample of $\mathcal{F}(\widetilde{\mathbb{M}})$. Its elements can be computed numerically in the following way: Let $X_\tau(x_0, \omega)$ denote the end point of the time-τ realization of X_t starting in $x_0 \in \mathbb{X}$ and outcome $\omega \in \Omega$, where Ω is the sample space underlying the process X_t. For $x \in \mathbb{X}$, $\tau > 0$ fixed as in Definition 2.2 and arbitrarily chosen $\{\omega_1, \ldots, \omega_M\} \subset \Omega$, let $y^{(k)}(x) := X_\tau(x, \omega_k)$. In short, the $y^{(k)}(x)$, $k = 1, \ldots, n$, sample the density p^t_x. In practice, the $y^{(k)}(x)$ will be generated

by multiple runs of a numerical SDE solver starting in x with M different random seeds ("bursts of simulations").

With the samples $y^{(k)}(x)$, we approximate $\mathcal{F}(p_x^t)$ by its Monte Carlo estimator:

$$\mathcal{F}(p_x^t) = \int \eta(x')\, p_x^t(x')\, \mathrm{d}x' \approx \underbrace{\frac{1}{M}\sum_{k=1}^{M} \eta(y^{(k)}(x))}_{=:\,\widehat{\mathcal{E}}(x)}.$$

Due to Proposition 2.7, the point cloud $\mathcal{F}\big(\{p_x^t \mid x \in \mathbb{X}_N\}\big)$, and for a large enough burst size M also its empirical estimator $\widehat{\mathcal{E}}\big(\mathbb{X}_N\big)$, then clusters around the r-dimensional manifold $\mathcal{F}(\mathbb{M})$ in $\mathbb{R}^{2r+1}$.

Parameterizing $\widehat{\mathcal{E}}\big(\mathbb{X}_N\big)$, i.e., finding the dominant nonlinear directions in this point cloud in $\mathbb{R}^{2r+1}$, now can be accomplished by a variety of classical manifold learning methods. We assume that we have a method at our disposal that is able to discover the underlying r-dimensional manifold within the point cloud $\widehat{\mathcal{E}}\big(\mathbb{X}_N\big)$, and assign each of the points $\{\widehat{\mathcal{E}}(x) \mid x \in \mathbb{X}_N\}$ a value $\tilde{\gamma}\big(\widehat{\mathcal{E}}(x)\big) \in \mathbb{R}^r$ according to its position on that manifold. For examples of such algorithms see Sect. 3.1. Hence, $\tilde{\gamma} : \widehat{\mathcal{E}}(\mathbb{X}_N) \to \mathbb{R}^k$ can be seen as an approximate parameterization of $\mathcal{F}(\mathbb{M})$, defined however only at the points $\widehat{\mathcal{E}}(\mathbb{X}_N)$. Any parameterization of $\mathcal{F}(\mathbb{M})$ in turn corresponds to a parameterization of $\mathbb{M}$, due to $\mathcal{F}$ being an embedding. Finally, any parameterization of $\mathbb{M}$ corresponds to a good reaction coordinate due to Theorem 2.5. Thus, the map $\xi(x) \colon \mathbb{X}_N \to \mathbb{R}^r$,

$$\xi(x) := \tilde{\gamma}\big(\widehat{\mathcal{E}}(x)\big),$$

forms a good reaction coordinate. Note however that it is only defined on the sample points $\mathbb{X}_N$.

The strategy of computing reaction coordinates by embedding densities sampled from $\widetilde{\mathbb{M}}$ into $\mathbb{R}^{2r+1}$ by a random linear map and learning a parameterization of the embedded manifold was first presented in Bittracher et al. (2017). The following algorithm summarizes the overall procedure:

Algorithm 2.1 Reaction coordinate computation based on Whitney embeddings.

Input: Transition manifold dimension r, intermediate lag time τ, matrix distribution σ.

1: Choose test points $\mathbb{X}_N = \{x_1, \ldots, x_N\}$ that cover the relevant parts of state space.
2: Randomly draw a matrix $A \in \mathbb{R}^{(2r+1)\times d}$ from σ. Define the map $\eta \colon x \mapsto Ax$.
3: **for** $i = 1, \ldots, N$ **do**
4: **for** $l = 1, \ldots, M$ **do**
5: Simulate trajectory of length τ with new random seed. Let the end point be denoted by $y_i^{(l)}$.
6: **end for**
7: **end for**
8: Compute the embedded empirical densities as $z_i \leftarrow \frac{1}{M}\sum_{j=1}^{M} \eta\big(y_i^{(j)}\big) \in \mathbb{R}^{2r+1}$.
9: Apply a nonlinear manifold learning algorithm to $\{z_i \mid i = 1, \ldots, N\}$. Let $\tilde{\gamma}(z_i) \in \mathbb{R}^r$ denote the resulting parametrization of the embedded test points.

Output: An r-dimensional reaction coordinate evaluated at the test points:

$$\xi(x_i) := \tilde{\gamma}(z_i), \quad i = 1, \ldots, N.$$

3 Kernel-Based Parameterization of the Transition Manifold

The approach described above for learning a parameterization of the transition manifold $\mathbb{M}$ by embedding it into Euclidean spaces requires a priori knowledge of the dimension of $\mathbb{M}$. Also, more importantly, $\mathbb{M}$ might be strongly distorted by the embedding $\mathcal{F}$, as described in Sect. 2.4. The kernel-based parameterization, which is the main novelty of this work, will address both of these shortcomings by embedding $\mathbb{M}$ into reproducing kernel Hilbert spaces.

3.1 Kernel Reformulation of the Embedding Algorithm

Manifold learning algorithms that can be used in Algorithm 2.1 include diffusion maps (Coifman et al. 2008), multidimensional scaling (Young 2013; Kruskal 1964), and locally linear embedding (Roweis and Saul 2000). These, and many others, require only a notion of distance between pairs of data points. In our case, this amounts to the Euclidean distances between embedded points, i.e., $\|\mathcal{E}(x_i) - \mathcal{E}(x_j)\|_2$, which can be computed by the Euclidean inner products $\langle \mathcal{E}(x_i), \mathcal{E}(x_j) \rangle$, as

$$\|\mathcal{E}(x_i) - \mathcal{E}(x_j)\|_2^2 = \langle \mathcal{E}(x_i), \mathcal{E}(x_i) \rangle - 2\langle \mathcal{E}(x_i), \mathcal{E}(x_j) \rangle + \langle \mathcal{E}(x_j), \mathcal{E}(x_j) \rangle.$$

Other compatible algorithms such as principal component analysis are based directly on the inner products. The inner products can be written as

$$\langle \mathcal{E}(x_i), \mathcal{E}(x_j) \rangle = \iint \langle \eta(y_i), \eta(y_j) \rangle \, p^t_{x_i}(y_i) \, p^t_{x_j}(y_j) \, \mathrm{d}y_i \, \mathrm{d}y_j,$$

and the empirical counterpart is

$$\langle \widehat{\mathcal{E}}(x_i), \widehat{\mathcal{E}}(x_j) \rangle = \frac{1}{M^2} \sum_{l_1, l_2 = 1}^{M} \langle \eta(y_i^{(l_1)}), \eta(y_j^{(l_2)}) \rangle.$$

However, rather than *explicitly* computing the inner product between the features on the right-hand side, we now assume that it can be computed *implicitly* by using a *kernel function* $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, i.e.,

$$k(y_i, y_j) = \langle \eta(y_i), \eta(y_j) \rangle. \tag{11}$$

That is, the previously randomly chosen linear observables η are now replaced by the feature mapping associated with the kernel function. This assumption, called the *kernel trick*, is commonly used to avoid the costly computation of inner products between high-dimensional features. However, instead of defining the kernel k based on previously chosen features, one typically considers kernels that implicitly define high- and possibly infinite-dimensional feature spaces. In this way, we are able to avoid the choice of the feature map η altogether.

Kernels with this property span a so-called *reproducing kernel Hilbert space*:

Definition 3.1 [Reproducing kernel Hilbert space (Schölkopf and Smola 2001)]. Let $k\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ be a positive definite function. A Hilbert space $\mathbb{H}$ of functions $f\colon \mathbb{X}\to\mathbb{R}$, together with the corresponding inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$ and norm $\|\cdot\|_{\mathbb{H}}=\sqrt{\langle\cdot,\cdot\rangle_{\mathbb{H}}}$ which fulfills

1. $\mathbb{H}=\overline{\operatorname{span}\{k(x,\cdot)\mid x\in\mathbb{X}\}}$, and
2. $\langle f, k(x,\cdot)\rangle_{\mathbb{H}}=f(x)$ for all $f\in\mathbb{H}$

is called the *reproducing kernel Hilbert space (RKHS)* associated with the *kernel* k.

Here, $\overline{A}$ denotes the completion of a set A with respect to $\|\cdot\|_{\mathbb{H}}$. Requirement 2 implies that

$$\langle k(x,\cdot),k(x',\cdot)\rangle_{\mathbb{H}}=k(x,x')\quad\text{for all } x,x'\in\mathbb{X}. \tag{12}$$

The inner product between general functions $f,g\in\operatorname{span}\{k(x,\cdot)\mid x\in\mathbb{X}\}$ can therefore be expressed as the weighted sum of kernel evaluations: Let

$$f=\sum_i \alpha_i k(x_i,\cdot),\quad g=\sum_j \beta_j k(x'_j,\cdot),$$

where the selection of points x_i, x'_j depends on f and g, respectively. Then

$$\langle f,g\rangle_{\mathbb{H}}=\sum_{i,j}\alpha_i\beta_j k(x_i,x'_j).$$

For functions on the boundary of $\operatorname{span}\{k(x,\cdot)\mid x\in\mathbb{X}\}$, the inner product is constructed by the usual limit procedure.

The map $\eta\colon x\mapsto k(x,\cdot)$ can be regarded as a function-valued feature map (the so-called *canonical feature map*). However, each positive definite kernel is guaranteed to also possess a feature map of at most countable dimension:

Theorem 3.2 [Mercer's theorem (Mercer 1909)]. *Let k be a positive definite kernel and ν be a finite Borel measure with support $\mathbb{X}$. Define the integral operator $\mathcal{T}_k\colon L^2_\nu\to L^2_\nu$ by*

$$\mathcal{T}_k f=\int k(\cdot,x)f(x)\,d\nu(x). \tag{13}$$

Then there is an orthonormal basis $\{\sqrt{\lambda_i}\varphi_i\}$ of $\mathbb{H}$ consisting of eigenfunctions φ_i of $\mathcal{T}_k$ rescaled with the square root of the corresponding nonnegative eigenvalues λ_i such that

$$k(x,x')=\sum_{i=0}^{\infty}\lambda_i\varphi_i(x)\varphi_i(x')\quad\textit{for all } x,x'\in\mathbb{X}. \tag{14}$$

The above formulation of Mercer's theorem has been taken from Muandet et al. (2017). The *Mercer features* $\eta_i:=\sqrt{\lambda_i}\varphi_i$ thus fulfill (11) for their corresponding kernel. The usage of the same symbol η as for the linear feature map from Sect. 2.4 is no coincidence, as the Mercer features will again serve the purpose to observe certain features of the full system. In what follows, $\eta(x)$ will always refer to the vector (or ℓ^2 sequence) defined by the Mercer features. If not stated otherwise, ν will be the standard Lebesgue measure.

Example 3.3 Examples of commonly used kernels are:

1. Linear kernel: $k(x, x') = x^\top x'$. One sees immediately that (11) is fulfilled by choosing $\eta_i(x) = x_i$, $i = 1, \ldots, n$ (also spanning the Mercer feature space).
2. Polynomial kernel of degree p: $k(x, x') = (x^\top x' + 1)^p$. It can be shown that the Mercer feature space is spanned by the monomials in x up to degree p.
3. Gaussian kernel: $k(x, x') = \exp\left(-\frac{1}{\sigma}\left\|x - x'\right\|_2^2\right)$, where $\sigma > 0$ is called the *bandwidth* of the kernel. Let $p \in \mathbb{N}$ and $\mathbf{p} = (p_1, \ldots, p_n)$ with $p_1 + \ldots + p_n = p$ be a multi-index. The Mercer features of k then take the form

$$\eta_{\mathbf{p}}(x) = e_{p_1}(x_1) \cdots e_{p_n}(x_n),$$

see Steinwart and Christmann (2008), where

$$e_{p_i}(x_i) = \sqrt{\frac{2^{p_i}}{\sigma^{2p_i} p_i!}}\, x_i^{p_i} \exp\left(-\tfrac{1}{\sigma^2} x_i^2\right). \qquad \triangle$$

Let $\mathcal{F}_k$ denote the density embedding based on the Mercer features of the kernel k, i.e.,

$$\left(\mathcal{F}_k(p_x^t)\right)_i := \int \eta_i(x') p_x^t(x') \mathrm{d}x', \quad i = 0, 1, 2, \ldots, \tag{15}$$

and let $\mathcal{E}_k(x) := \mathcal{F}_k(p_x^t)$. The amount of information about p_x^t preserved by the embedding $\mathcal{F}_k$ depends on the choice of the kernel k. For the first two kernels in Example 3.3, the information preserved has a familiar stochastic interpretation (see, e.g., Muandet et al. 2017; Schölkopf et al. 2015; Sriperumbudur et al. 2010):

1. Let k be the linear kernel. Then

$$\left\|\mathcal{F}_k(p_{x_1}^t) - \mathcal{F}_k(p_{x_2}^t)\right\|_2 = 0 \iff \int p_{x_1}^t(y)\mathrm{d}y = \int p_{x_2}^t(y)\mathrm{d}y,$$

   i.e., the means of $p_{x_1}^t$ and $p_{x_1}^t$ coincide.
2. Let k be the polynomial kernel of degree $p > 1$. Then

$$\left\|\mathcal{F}_k(p_{x_1}^t) - \mathcal{F}_k(p_{x_2}^t)\right\|_2 = 0 \iff \mathbf{m}_i\left(p_{x_1}^t\right) = \mathbf{m}_i\left(p_{x_2}^t\right), \quad i = 1, \ldots, p,$$

   i.e., the first p moments $\mathbf{m}_i$ of $p_{x_1}^t$ and $p_{x_1}^t$ coincide.

Remark 3.4 In practice, comparing the first p moments often is enough to sufficiently distinguish the transition densities that constitute the transition manifold. However, densities that differ only in higher moments cannot be distinguished by $\mathcal{F}_k$, which means that for the above two kernels, $\mathcal{F}_k$ is not injective on $\mathbb{M}$. Therefore, $\mathcal{F}_k$ does not belong to the prevalent class of maps that is at the heart of the Whitney embedding theorem 2.6. We can therefore not utilize the Whitney embedding theorem to argue that the topology of $\mathbb{M}$ is preserved under $\mathcal{F}_k$. Instead, in Sect. 3.2, we will use a different argument to show that the embedding is indeed injective for the Gaussian kernel (and others).

Still, by formally using the Mercer dynamical embedding $\mathcal{E}_k$ in (10) (abusing notation if there are countably infinitely many such features), and using the kernel trick, we can now reformulate Algorithm 2.1 as a *kernel-based* method that does not require the explicit computation of any feature vector. This is summarized in Algorithm 3.1.

Algorithm 3.1 Kernel-based computation of the reaction coordinate.

Input: Kernel $k\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$, intermediate lag time τ.

1: Choose test points $\mathbb{X}_N = \{x_1,\dots,x_N\}$ that cover the relevant parts of state space.
2: **for** $i = 1,\dots,N$ **do**
3: **for** $l = 1,\dots,M$ **do**
4: Simulate trajectory of length τ with new random seed. Let the end point be denoted by $y_i^{(l)}$.
5: **end for**
6: **end for**
7: Compute the kernel matrix $K\in\mathbb{R}^{N\times N}$:

$$K_{ij} = \frac{1}{M^2}\sum_{l_1,l_2=1}^{M} k(y_i^{(l_1)}, y_j^{(l_2)}).$$

8: Compute the distance matrix $D\in\mathbb{R}^{N\times N}$:

$$D_{ij} = K_{ii} + K_{jj} - 2K_{ij}.$$

9: Apply a distance-based manifold learning algorithm to the distance matrix D. Denote the resulting parametrization of the underlying i-th element by $\widetilde{\gamma}_i\in\mathbb{R}^r$.

Output: An r-dimensional reaction coordinate evaluated at the test points:

$$\xi(x_i) := \widetilde{\gamma}_i,\quad i = 1,\dots,N.$$

3.2 Condition Number of the Kernel Embedding

We will now investigate to what extent the kernel embedding preserves the topology and geometry of the transition manifold.

3.2.1 Kernel Mean Embedding

We derived the kernel-based algorithm by considering the embedding $\mathcal{F}_k$ of the transition manifold into the image space of the Mercer features in order to highlight the similarity to the Whitey embedding based on randomly drawn features. Of course, the Mercer features never had to be computed explicitly.

However, in order to investigate the quality of this embedding procedure, it is advantageous to consider a different, yet equivalent embedding map: The transition manifold can be directly embedded into the RKHS by means of the *kernel mean embedding* operator.

Definition 3.5 Let k be a positive definite kernel and $\mathbb{H}$ the associated RKHS. Let p be a probability density over $\mathbb{X}$. Define the *kernel mean embedding* of p by

$$\mu(p) := \int_{\mathbb{X}} k(x,\cdot)\,p(x)\,\mathrm{d}x$$

and the *empirical kernel mean embedding* by

$$\hat{\mu}(p) := \frac{1}{m} \sum_i k(x_i, \cdot) \quad \text{with} \quad \{x_1, \ldots, x_m\} \sim p.$$

Note that $\mu(p)$ and $\hat{\mu}(p)$ are again elements of $\mathbb{H}$ and that for ν in (13) being the Lebesgue measure we obtain $\mu(p) = \mathcal{T}_k\, p$. Further, one sees that

$$\left\langle \mathcal{F}_k(p^t_{x_1}), \mathcal{F}_k(p^t_{x_2}) \right\rangle = \left\langle \mu(p^t_{x_1}), \mu(p^t_{x_2}) \right\rangle_{\mathbb{H}},$$

where the inner product $\langle \cdot, \cdot \rangle$ refers to the Euclidean inner product or the inner product in $\ell^2(\mathbb{N}_0)$, dependent on whether $\mathcal{F}_k(p)$ is finite or countably infinite. Thus, for investigating whether the embedding $\mathcal{F}_k$ preserves distances or inner products between densities, we can equivalently investigate the embedding μ. This is advantageous as injectivity and isometry properties of the kernel mean embedding are well-studied.

3.2.2 Injectivity of the Kernel Mean Embedding

A first important result is that k can be chosen such that μ is injective. Such kernels are called *characteristic* (Fukumizu et al. 2007). In Sriperumbudur et al. (2010), several conditions for characteristic kernels are listed, including the following:

Theorem 3.6 (Sriperumbudur et al. 2010, Theorem 7). *The kernel k is characteristic if for all $f \in L_2,\ f \neq 0$ it holds that*

$$\int_{\mathbb{X}} \int_{\mathbb{X}} k(x, x') f(x) f(x') \mathrm{d}x \, \mathrm{d}x' > 0. \tag{16}$$

Condition (16) is known as the *Mercer condition*, which is, for example, fulfilled by the Gaussian kernel from Example 3.3. The Mercer features of such a kernel are particularly rich.

Theorem 3.7 *Assume that the kernel satisfies the Mercer condition* (16)*. Then the eigenfunctions $\{\psi_i\}$ of $\mathcal{T}_k$ form an orthonormal basis of $L^2(\nu)$.*

For more details, see, e.g., Schölkopf and Smola (2001) and Steinwart and Christmann (2008). It is easy to see that for kernels fulfilling (16), μ as a map from L^2 to $\mathbb{H}$ is Lipschitz continuous:

Lemma 3.8 *Let k be a characteristic kernel with Mercer eigenvalues λ_i, $i \in \mathbb{N}_0$. Then $\mu\colon L^2 \to \mathbb{H}$ is Lipschitz continuous with constant*

$$c := \sqrt{\lambda_0}. \tag{17}$$

Proof As μ is linear, it suffices to show that $\|\mu(f)\|_{\mathbb{H}} \leq c\|f\|_2$ for all $f \in L_2$. We obtain

$$\begin{aligned}\|\mu(f)\|_{\mathbb{H}}^2 &= \langle \mu(f), \mu(f)\rangle_{\mathbb{H}} \\ &= \left\langle \int_{\mathbb{X}} f(x)k(x,\cdot)\,\mathrm{d}x, \int_{\mathbb{X}} f(x)k(x,\cdot)\,\mathrm{d}x \right\rangle_{\mathbb{H}} \\ &= \int_{\mathbb{X}}\int_{\mathbb{X}} f(x)f(y)k(x,y)\,\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

where (12) was used in the last line. By expanding k into its Mercer features via (14), this becomes

$$\begin{aligned}\|\mu(f)\|_{\mathbb{H}}^2 &= \int_{\mathbb{X}}\int_{\mathbb{X}} f(x)f(y)\left(\sum_{i\in\mathbb{N}} \lambda_i \varphi_i(x)\varphi_i(y)\right)\mathrm{d}x\,\mathrm{d}y \\ &= \sum_{i\in\mathbb{N}} \lambda_i \langle f, \varphi_i\rangle_{L^2}^2.\end{aligned}$$

By Theorem 3.7, the φ_i form an orthonormal basis of L^2, and thus

$$\|\mu(f)\|_{\mathbb{H}}^2 \leq \lambda_0 \|f\|_{L^2}^2.$$

□

Thus, if the kernel is characteristic, the structure of the TM and the fuzzy TM are qualitatively preserved under the embedding.

Corollary 3.9 *Let k be a characteristic kernel and let X_t be (ε, r)-reducible. Then $\mu(\mathbb{M}) \subset \mathbb{H}$ has an r-dimensional parameterization, and for all $x \in \mathbb{X}$ it holds that*

$$\inf_{g\in\mu(\mathbb{M})} \|g - \mu(p_x^t)\|_{\mathbb{H}} \leq \sqrt{\lambda_0}\|\sqrt{\rho}\|_\infty \varepsilon.$$

Proof By Lemma 3.8, the map $\mu\colon L^2 \to \mathbb{H}$ is Lipschitz continuous (and furthermore injective), and thus any r-dimensional (local) parameterization $\chi : \Omega \subset \mathbb{R}^r \to L^2_{1/\rho}$ of $\mathbb{M}$ yields an r-dimensional parameterization $\mu \circ \chi$ of $\mu(\mathbb{M})$. For $x \in \mathbb{X}$, consider now any $f \in L^2_{1/\rho}$ with $\|f - p_x^t\|_{L^2_{1/\rho}} \leq \varepsilon$. For $g := \mu(f)$, we then get

$$\left\|g - \mu(p_x^t)\right\|_{\mathbb{H}} = \left\|\mu\left(f - p_x^t\right)\right\|_{\mathbb{H}}$$

which by Lemma 3.8 is

$$\begin{aligned}&\leq \sqrt{\lambda_0}\|\sqrt{\rho}\|_\infty \|f - p_x^t\|_{L^2_{1/\rho}} \\ &\leq \sqrt{\lambda_0}\|\sqrt{\rho}\|_\infty \varepsilon.\end{aligned}$$

□

Remark 3.10 This result should be seen as an analogue to Proposition 2.7 for the Whitney-based TM embedding. In short, for characteristic kernels, the injectivity and continuity of μ guarantee that the image of $\mathbb{M}$ under μ is again an r-dimensional object in $\mathbb{H}$, and Corollary 3.9 guarantees that the embedded fuzzy transition manifold $\mu(\widetilde{\mathbb{M}})$ still clusters closely around $\mu(\mathbb{M})$ (if $\sqrt{\lambda_0}$ and $\|\sqrt{\rho}\|_\infty$ in Corollary 3.9 are small). This again guarantees a minimal degree of well-posedness of the problem.

3.2.3 Distortion Under the Kernel Mean Embedding

Unlike the Whitney embedding, the kernel embedding now allows us to derive conditions under which the distortion of $\mathbb{M}$ is bounded. We have to show that the $L^2_{1/\rho}$-distance between points on $\mathbb{M}$ is not overly decreased or increased by the kernel mean embedding. To formalize this, we consider measures for the manifold's internal distortion, following the notions of metric embedding theory (Abraham et al. 2011). We call the embedding *well-conditioned* if both the

$$\text{contraction: } \sup_{\substack{p,q\in\mathbb{M}\\ q\neq p}} \frac{\|p-q\|_{L^2_{1/\rho}}}{\|\mu(p)-\mu(q)\|_{\mathbb{H}}} \quad \text{and the expansion: } \sup_{\substack{p,q\in\mathbb{M}\\ q\neq p}} \frac{\|\mu(p)-\mu(q)\|_{\mathbb{H}}}{\|p-q\|_{L^2_{1/\rho}}} \tag{18}$$

are small (close to one). Here, μ denotes the embedding corresponding to a characteristic kernel.

Due to the Lipschitz continuity of μ (see Lemma 3.8) and $\|\cdot\|_{L^2} \leq \|\sqrt{\rho}\|_\infty \|\cdot\|_{L^2_{1/\rho}}$, we have

$$\frac{\|\mu(p)-\mu(q)\|_{\mathbb{H}}}{\|p-q\|_{L^2_{1/\rho}}} \leq \sqrt{\lambda_0}\|\sqrt{\rho}\|_\infty, \tag{19}$$

thus bounding the expansion.

Contraction bound: regularity requirement Unfortunately, it is not possible even for characteristic kernels to derive a bound for the contraction that holds uniformly for all $p, q \in L^2_{1/\rho}$, as the following proposition shows. Nevertheless, we will be able to give reasonable bounds under some regularity- and dynamic assumptions, (21) and (24), respectively.

Proposition 3.11 (Unbounded inverse embedding). *Assume the kernel embedding operator μ has absolutely bounded orthonormal eigenfunctions φ_i with corresponding nonnegative eigenvalues λ_i (arranged in nonincreasing order). Assume* $\lim_{i\to\infty}\lambda_i = 0$. *Then, there exist functions $p, q \in L^2_{1/\rho}$ such that*

$$\frac{\|p-q\|_{L^2_{1/\rho}}}{\|\mu(p)-\mu(q)\|_{\mathbb{H}}} > \frac{1}{\varepsilon}$$

for any arbitrarily small $\varepsilon > 0$.

Proof See "Appendix A". □

The assumptions of Proposition 3.11 are fulfilled for example for the Gaussian kernel. A similar but non-quantitative result has been derived in Sriperumbudur et al. 2010, Theorem 19. The idea behind its proof and the proof of Proposition 3.11 is that, if p and q vary only in higher eigenfunctions φ_i of the embedding operator μ (see also Theorem 3.2), the $\mathbb{H}$-distance can become arbitrarily small. If, however, we can reasonably restrict our considerations to the subclass of functions whose variation in the higher φ_i is small compared to the variation in the lower φ_i, a favorable bound can be derived. Let the expansion of $h = p - q$ be given by

$$h = \sum_{i=0}^{\infty} \tilde{h}_i \varphi_i$$

with the sequence $(\tilde{h}_0, \tilde{h}_1, \ldots) \in \ell^2$. Now, for any $i_{\max} \in \mathbb{N}$ such that there exists an index $i \leq i_{\max}$ with $\tilde{h}_i \neq 0$, define the factor

$$c(h, i_{\max}) := 1 + \frac{\sum_{i=i_{\max}+1}^{\infty} \tilde{h}_i^2}{\sum_{i=0}^{i_{\max}} \tilde{h}_i^2}. \tag{20}$$

This factor bounds the contribution of the higher Mercer eigenfunctions to h by the contribution of the lower ones, hence it is a *regularity bound*:

$$\sum_{i=0}^{\infty} \tilde{h}_i^2 = c(h, i_{\max}) \cdot \sum_{i=0}^{i_{\max}} \tilde{h}_i^2.$$

Thus, for an individual h, we can bound the distortion of the L^2-norm under μ with the help of $c(h, i_{\max})$.

Lemma 3.12 *Let $h \in L^2$, $i_{max} \in \mathbb{N}$, and $c(h, i_{max})$ be defined as in (20). Then*

$$\|\mu(h)\|_{\mathbb{H}} \geq \sqrt{\frac{\lambda_{i_{max}}}{c(h, i_{max})}} \|h\|_{L^2}.$$

Proof See "Appendix A". □

We from now on make the assumption that for every index $i_{\max}$ there exists a constant $c^*_{i_{\max}} > 0$ such that

$$c(p^{\tau}_{x_1} - p^{\tau}_{x_2}, i_{\max}) \leq c^*_{i_{\max}} \tag{21}$$

for all $p^{\tau}_{x_1}, p^{\tau}_{x_2} \in \mathbb{M}$. The existence and form of this constant strongly depends on the shape of the Mercer eigenfunctions, hence the kernel. However, we motivate the existence of such a global constant by the observation that higher Mercer eigenfunctions typically consist of high Fourier modes, and that these modes decay quickly under the

dynamics. Therefore, high Mercer eigenfunctions should have a negligible share of the p_x^τ and the differences $p_{x_1}^\tau - p_{x_2}^\tau$. For such $p_{x_1}^\tau$, $p_{x_2}^\tau$, we thus have

$$\|\mu(p_{x_1}^\tau) - \mu(p_{x_2}^\tau)\|_{\mathbb{H}} \geq \sqrt{\frac{\lambda_{i_{\max}}}{c^*_{i_{\max}}}} \|p_{x_1}^\tau - p_{x_2}^\tau\|_{L^2}. \tag{22}$$

Contraction bound: dynamical requirements Note that (22) is only an intermediate step for deriving a contraction bound, as the relevant distance measure in Definition 2.4 is the $L^2_{1/\rho}$-norm, for reasons detailed in Sect. 2.2, and (22) measures the density distance in the L^2-norm. Unfortunately, a naive estimation yields

$$\|h\|_{L^2_{1/\rho}} \leq \left\|1/\sqrt{\rho}\right\|_\infty \|h\|_{L^2}. \tag{23}$$

While due to ergodicity $1/\sqrt{\rho}$ is indeed defined on all of $\mathbb{X}$, it becomes large in regions of small invariant measure ρ, i.e., "dynamically irrelevant" regions. This would lead to a very large upper bound for the contraction. For general h, a more favorable estimate is indeed difficult to obtain. For us, however, $h = p_{x_1}^\tau - p_{x_2}^\tau$, and we can utilize that these "dynamically irrelevant" regions are almost never visited by the system.

To formalize this, we require one additional assumption that can be justified by the metastability of the system. One defining characteristic of metastable systems is the phenomenon that essentially any trajectory moves *nearly instantaneously*[3] into one of the metastable sets before continuing. With $A_i, \ldots, A_d \subset \mathbb{X}$ denoting these sets, we can thus assume that the probability density $p_x^\tau(y)$ to move from x to y in time τ depends almost only on the probabilities to (instantaneously) move to the sets A_i from x (denoted by $c_i(x)$, thus $\sum_{i=1}^d c_i(x) \leq 1$) and the probabilities to then move from A_i to y in time τ (denoted by $p_{A_i}^\tau(y)$), i.e.,

$$p_x^\tau \approx \sum_{i=1}^d c_i(x) p_{A_i}^\tau.$$

To be more precise, we require for all $x, y \in \mathbb{X}$ that

$$\frac{\left\|p_x^\tau - \sum_{i=1}^d c_i(x) p_{A_i}^\tau\right\|_\infty}{\|p_x^\tau\|_\infty} \approx 0,$$

which is equivalent to

$$p_x^\tau(y) = \frac{1}{1-\delta(x,y)} \sum_{i=1}^d c_i(x) p_{A_i}^\tau(y) \tag{24}$$

[3] Here, "nearly instantaneously" is to be understood in the sense that there is an "attraction time" $\tau_a \ll \tau$ such that starting from essentially any initial condition the system will enter one of the metastable sets within time τ_a with overwhelming probability. Thus, choosing τ as an intermediate time that is larger than the non-metastable time scales of local fluctuations, is essential. However, as we discuss in Remark 3.15, it is also imperative not to choose τ too large.

for some positive function $\delta\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ with $\|\delta\|_\infty \le \delta^* \ll 1$. The positivity of δ comes from the fact that there is a miniscule, but positive probability (density) to move from x to y without first equilibrating inside a metastable set, thus $p_x^\tau(y) > \sum_{i=1}^d c_i(x)p_{A_i}^\tau(y)$.

With this, we can bound the invariant density ρ from below as follows:

$$\begin{aligned}\rho(y) &= \int_{\mathbb{X}} \rho(x)p_x^\tau(y)\,\mathrm{d}x\\ &= \int_{\mathbb{X}} \rho(x)\frac{1}{1-\delta(x,y)}\sum_{i=1}^d c_i(x)p_{A_i}^\tau(y)\,\mathrm{d}x\\ &> \int_{\mathbb{X}} \rho(x)\sum_{i=1}^d c_i(x)p_{A_i}^\tau(y)\,\mathrm{d}x\\ &= \sum_{i=1}^d \underbrace{\left(\int_{\mathbb{X}}\rho(x)c_i(x)\,\mathrm{d}x\right)}_{=:b_i} p_{A_i}^\tau(y). \end{aligned} \tag{25}$$

The b_i can be seen as the equilibrium probability mass almost instantaneously attracted to A_i. For every important metastable set this will not be too small.

As a first step, (25) allows us to bound the $L^2_{1/\rho}$-norm of p_x^τ by their L^1-norm:

Lemma 3.13 *Let the assumption* (24) *hold and* b_i, $i = 1,\dots,d$, *be defined as in* (25). *Then for any* $x\in\mathbb{X}$ *it holds that*

$$\|p_x^\tau\|^2_{L^2_{1/\rho}} < \frac{1}{(1-\delta^*)\min_i b_i}\|p_x^\tau\|_{L^1} = \frac{1}{(1-\delta^*)\min_i b_i} \tag{26}$$

Proof See "Appendix A". □

This shows that indeed $p_x^\tau \in L^2_{1/\rho}$, as required by Definition 2.2. Further, Hölder's inequality gives

$$\|f\|_{L^1} = \int_{\mathbb{X}} |f(x)|\cdot 1\,\mathrm{d}x \le \|f\|_{L^2}\cdot|\mathbb{X}|^{1/2}.$$

where $|\mathbb{X}|$ denotes the Lebesgue measure of the state space. This now also allows us to bound the $L^2_{1/\rho}$-norm of the p_x^τ by their L^2-norm:

Lemma 3.14 *Let the assumption* (24) *hold and* b_i, $i = 1,\dots,d$, *be defined as in* (25). *Then for any* $x_1, x_2\in\mathbb{X}$ *it holds that*

$$\|p_{x_1}^\tau - p_{x_2}^\tau\|^2_{L^2_{1/\rho}} < \frac{|\mathbb{X}|^{1/2}}{(1-\delta^*)\min_i b_i}\|p_{x_1}^\tau - p_{x_2}^\tau\|_{L^2}. \tag{27}$$

Proof See "Appendix A". □

Of course, due to the squared norm on the left-hand side, this is not a Lipschitz bound. However, recall that our main motivation for deriving a bound for the contraction is to show that large distances in $L^2_{1/\rho}$ are not overly compressed under the embedding into $\mathbb{H}$, as illustrated in Fig. 2. We therefore abstain from deriving such a bound for very small distances in $L^2_{1/\rho}$ and only estimate the contraction of pairs of densities $p^\tau_{x_1}, p^\tau_{x_2}$ with

$$\|p^\tau_{x_1} - p^\tau_{x_2}\|_{L^2_{1/\rho}} \geq C \tag{28}$$

for some constant $C > 0$. That C is reasonably large is discussed in Remark 3.15 below. For such differences, we can then relate the L^2 to the $L^2_{1/\rho}$-norm, i.e.,

$$\|p^\tau_{x_1} - p^\tau_{x_2}\|_{L^2_{1/\rho}} < \frac{|\mathbb{X}|^{1/2}}{C(1-\delta^*)\min_i b_i}\|p^\tau_{x_1} - p^\tau_{x_2}\|_{L^2}.$$

Together with Lemma 3.12 and assumption (21), this gives

$$\|p^\tau_{x_1} - p^\tau_{x_2}\|_{L^2_{1/\rho}} < \frac{|\mathbb{X}|^{1/2}}{C(1-\delta^*)\min_i b_i}\sqrt{\frac{\lambda_{i_{\max}}}{c^*_{i_{\max}}}}\left\|\mu\left(p^\tau_{x_1}\right) - \mu\left(p^\tau_{x_2}\right)\right\|_{\mathbb{H}}, \tag{29}$$

which is our contraction bound.

Remark 3.15 For the distortion of the transition manifold under the embedding the essential property is that the global "spanning structure" of the manifold is well preserved. In other words, the embedded p^τ_x, p^τ_y should be well-separated for $x, y \in \mathbb{X}$ from different metastable sets A_i. Since the embedding is continuous, the transition paths connecting them will be preserved as well.

As $p^\tau_{A_i} \to \rho$ as $\tau \to \infty$ for every i, it is important that τ is not too large, such that the transition manifold is a meaningful object. In a other words, τ should be such that $p^\tau_{A_i}$ and $p^\tau_{A_j}$ are sufficiently distinct for $i \neq j$. Thus, we *require* that $\|p^\tau_{A_i} - p^\tau_{A_j}\|_{L^2_{1/\rho}}$ is sufficiently large for $i \neq j$, hence C can be chosen as a constant such that $1/C$ is reasonably small.

Remark 3.16 The bounds (19) and (29) guarantee the well-posedness of the overall embedding and parameterization problem as the relevant expansion and contraction of the transition manifold cannot become arbitrarily large. We will support this statement with numerical evidence (see Sect. 4.2) showing that the distortion is, in practice, indeed small. It should be noted, however, that we do not expect the analytically derived bounds to perform well as quantitative error estimates, as many of the estimates that led to them are rather rough.

4 Illustrative Examples and Applications

We now evaluate the performance of Algorithm 3.1 on several example systems. In Sect. 4.1, the reaction coordinate of the well-known Müller–Brown potential is computed. In Sect. 4.2, the maximum distortion of the transition manifold under the

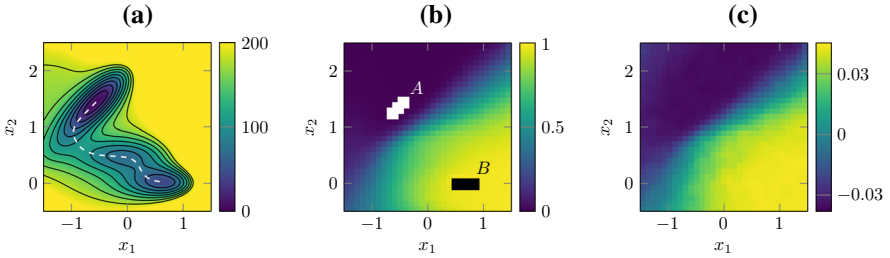

Fig. 3 **a** The Müller–Brown potential energy function with its three characteristic local minima and the connecting MEP (white line). **b** The committor function q_{AB} associated with the areas around the top left (A) and bottom right (B) energy minimum. **c** Reaction coordinate ξ of the Müller–Brown potential, computed by Algorithm 3.1

Whitney and kernel embedding with different parameters is quantified using a specifically constructed toy system. Finally, in Sect. 4.3, we demonstrate the applicability of our method to molecular dynamics problems.

The code used to generate our results is provided in the form of MATLAB scripts in the supplementary material. In the case of the Alanine dipeptide, the MD dataset is also provided.

4.1 Reaction Coordinate of the Müller–Brown Potential

As a first illustrating example, we compute the reaction coordinate of the two-dimensional Müller–Brown potential (Müller 1980) via the new kernel-based Algorithm 3.1. Originally a model for reaction energy profiles of chemical transformations, this potential has become a standard benchmark system for methods computing reaction coordinates and transition pathways of metastable systems as well as enhanced sampling techniques (Vanden-Eijnden and Venturoli 2009; Elber et al. 2017; Frewen et al. Oct. 2009).

The potential energy surface (see Fig. 3a) possesses three local minima, where the two bottom minima are separated only by a rather shallow energy barrier. Correspondingly, the system's characteristic long-term behavior is determined by the rare transitions between the minima. These transitions happen predominantly along the potential's *minimum energy pathway* (MEP), which is shown as white dashed line and was computed using the zero temperature string method (E et al. 2002, 2007).

For two sets $A, B \subset \mathbb{X}$ and a starting point $x \in \mathbb{X}$, the *committor function* $q_{AB}(x)$ is defined as the probability that the process hits set A before hitting set B, provided it started in x at time zero. For a precise definition see Schütte et al. (2013). For the Müller–Brown potential, the committor function associated with the top left and bottom right energy minima, shown in Fig. 3b, can be considered an optimal reaction coordinate (Elber et al. 2017). Therefore, we use the (qualitative) comparison with the committor function as a benchmark for our reaction coordinate. Note that the computation of the committor function requires global knowledge of the metastable sets and is often not a practical option for the identification of reaction coordinates.

The governing dynamics is given by an overdamped Langevin equation (3), which we solve numerically using the Euler–Maruyama scheme. At inverse temperature $\beta = 0.05$, eigenanalysis of the Perron-Frobenius operator reveals that the lag time $\tau = 0.03$ falls between the slow and fast time scales (see Prinz et al. (2011) for the technique to determine the time scales) and is thus chosen as the intermediate lag time[4]. The test points $\{x_1, \ldots, x_N\}$ required by Algorithm 3.1 are given by a regular 32×32 grid discretization of the domain $[-1.5, 1.5] \times [-0.5, 2.5]$. For the embedding, the Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{\sigma}\right) \tag{30}$$

with bandwidth $\sigma = 0.1$ is used, and subsequently the distance matrix D is computed.

Finally, for the manifold learning task in Algorithm 3.1, the diffusion maps algorithm with bandwidth parameter $\sigma' = 0.1$ is applied to D. In order to obtain the diffusion map coordinates, a Markov matrix based on the pairwise distances d_{ij} is constructed by computing

$$M_{ij} = \frac{K_{ij}}{s_i},$$

where $K_{ij} = \exp(-\frac{D_{ij}}{\sigma'})$ and $s_i = \sum_j K_{ij}$. Depending on some parameter that determines the approximated differential operator, different normalization steps might be involved. The diffusion map is then given by the eigenvectors of this matrix. Details pertaining to the technique's derivation and interpretation can be found in Nadler et al. (2006).

The reaction coordinate ξ for the test points is shown in Fig. 3c. We observe remarkable resemblance to the committor function.

4.1.1 On the Choice of the Kernel Parameters

The correct choice of the kernel parameters σ and σ' above is essential for the performance of our method. In particular, σ influences the eigenvalues λ_i of the integral operator (13), and hence the distortion of the TM under the embedding (see 29). For the Müller–Brown potential, the final RC is very insensitive to variations of σ, e.g., choosing $\sigma = 0.01$ or $\sigma = 1$ still leads to qualitatively very similar RCs. A systematic analysis of the influence of σ on the distortion for a different system is presented in the next section. However, the appropriate choice of the kernel for the embedding is a hard problem in general for kernel-based methods, and kernel optimization is an active field of research (Gönen and Alpaydin 2011; Duvenaud et al. 2013; Owhadi and Yoo 2019).

As such, also the choice of the diffusion maps kernel parameter σ' is nontrivial in general. Luckily, however, σ' can be optimized *after* the computationally hard part of Algorithm 3.1 (the computation of D), and there exist stand-alone optimization methods for this purpose (Gaspar et al. 2012; Berry and Harlim 2016; Lee and Verleysen 2009). In any case, the diffusion maps algorithm is not an intrinsic part of

[4] In realistic applications such as molecular dynamics, computation of the Perron-Frobenius operator spectrum may be numerically infeasible. Here, certain heuristics may be available to estimate τ, see Sect. 4.3.

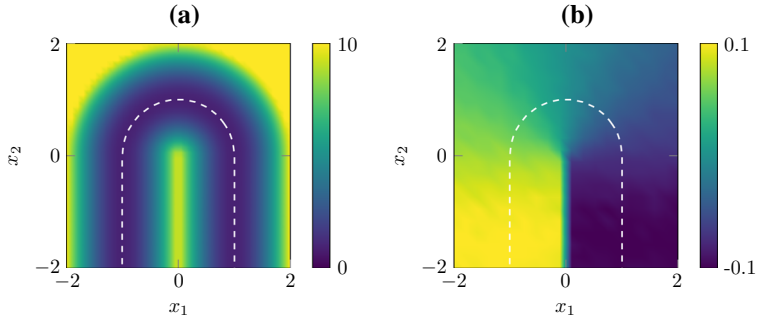

Fig. 4 Horseshoe potential. **a** Potential energy function. **b** Reaction coordinate computed with the kernel algorithm. The white dashed line represents the MEP

Algorithm 3.1 and can in principle be replaced by some manifold learning method that does not require parameter optimization.

We would also like to point out that the kernel evaluations used for the RKHS embedding of densities and the kernel evaluations used in the diffusion maps algorithm should not be mixed up as they serve entirely different purposes. The former is used to embed the state space densities into $\mathbb{H}$, while the latter is used to approximate the Laplace–Beltrami operator on the manifold in $\mathbb{H}$ that is to learn (this is the principle on which the diffusion maps algorithm is based). Even though the Gaussian kernel is a popular choice due to its favorable characteristics, one has great freedom in choosing a kernel for the RKHS-embedding, whereas in the classical diffusion maps algorithm, predominantly the Gaussian kernel is used, so the repeated use of the Gaussian kernel does not constitute a connection. Moreover, the fact that in this example identical bandwidth parameters were used was a mere coincidence. We do not see a way to unify these kernel evaluations, neither on a conceptual nor algorithmic level.

4.2 Distortion Under the Whitney and Kernel Embeddings

We now demonstrate the distortion of the fuzzy transition manifold under the embedding via the conventional Algorithm 2.1 (Whitney embedding) and how our new kernel-based Algorithm 3.1 is able to reduce that distortion. To this end, we consider the two-dimensional potential depicted in Fig. 4a. This potential is particularly suited for demonstrating the distortion, as the parts of the transition manifold that correspond to the two parallel "branches" of the MEP easily are mapped very close to each other when choosing a "bad" embedding of form (9) (cf. also Fig. 2).

We again consider the diffusion process (3) in this potential, at inverse temperature $\beta = 2$. The MEP of this potential is shown as a white dashed line in Fig. 4a. Individual trajectories equilibrate quickly perpendicular to the MEP (but not accross the central vertical barrier), and equilibrate slowly along the MEP. Hence, a good RC should parameterize the MEP, and stay constant perpendicular to it. For demonstration purposes, such a reaction coordinate is depicted in Fig. 4b.

Our aim here, however, is to estimate the distortion of the fuzzy TM $\widetilde{\mathbb{M}}$ under various embeddings based on a finite number of samples of $\widetilde{\mathbb{M}}$. The basic procedure is the following: We draw $N = 200$ test points x_i uniformly randomly from the region $\mathbb{X} = [-2, 2]^2$ to cover the state space evenly. We then estimate the densities $p^{\tau}_{x_i}$ by M simulations of length τ (Monte Carlo sampling). Here $\tau = 1$ was chosen again based on the eigenanalysis of the Perron–Frobenius operator. For $M = 1000$, we consider the Monte Carlo sampling sufficiently converged. Finally, we embed the estimated $p^{\tau}_{x_i}$ via Whitney and kernel methods with various parameters, compute pairwise distances between the embeddings, and compare it to the L^2_ρ-distances between the $p^{\tau}_{x_i}$.

4.2.1 Whitney Embedding

For the Whitney embedding, the expected manifold dimension $r = 1$ is assumed to be known in advance. To demonstrate the different effects of "good" and "bad" embedding functions, two $2r + 1$-dimensional linear observables $\eta : \mathbb{R}^2 \to \mathbb{R}^3$ were chosen:

$$\eta_g \colon x \mapsto A_g x, \qquad \eta_b \colon x \mapsto A_b x, \qquad A_g, A_b \in \mathbb{R}^{3\times 2},$$

and the corresponding embedding functions $\mathcal{F}_g, \mathcal{F}_b$ constructed via (9). The coefficients of A_g of the observable function η_g were chosen randomly via the MATLAB command

$$\texttt{rng(1)}; \texttt{A}_\texttt{g} = \texttt{rand(3, 2)} - \texttt{0.5},$$

which resulted in the matrix

$$A_g \approx \begin{pmatrix} -0.08 & -0.20 \\ 0.22 & -0.35 \\ -0.49 & -0.41 \end{pmatrix}.$$

Under the embedding $\mathcal{F}_g$, a "horseshoe-like" structure, corresponding to the embedded TM, can indeed be distinguished very well, see Fig. 5a. This is due to the fact that two distinct points of the MEP, in particular on the two opposite branches, are never mapped to the same point under η_g.

On the other hand, the matrix A_b of the "bad" observable function η_b was intentionally constructed to consist of three row vectors that are pairwise almost linearly dependent, and that essentially ignore the x_1-component of state space points:

$$A_b = \begin{pmatrix} 0 & 1 \\ \varepsilon & 1+\varepsilon \\ -\varepsilon & 1-\varepsilon \end{pmatrix} \qquad \text{with } \varepsilon = 0.05.$$

This way, points on the two opposite branches of the MEP but with the same x_2-coordinate are mapped to almost the same point in $\mathbb{R}^3$. The result is an embedding

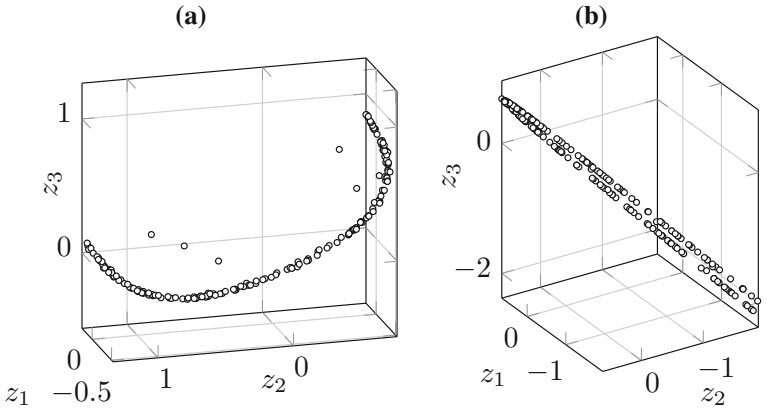

Fig. 5 Whitney embeddings of the test points for different observable functions. **a** Pairwise strongly linearly independent coefficient vectors, i.e., "good" observables. **b** Almost pairwise linearly dependent coefficient vectors, i.e., "bad" observables

of the TM in which the two branches can hardly be distinguished, see Fig. 5b. This would make the numerical identification of the manifold structure extremely difficult.

Note that the judgment of quality of the embedding function has to be performed manually *after* the embedding, as it is impossible to reliably choose good embedding functions without detailed a priori knowledge of the global structure of the transition manifold or transition pathway. While in our experience, randomly chosen coefficients typically result in "good-enough" embedding functions, this uncertainty in the numerical algorithm should be seen as one of the main reasons to use the more consistent kernel embeddings instead.

4.2.2 Kernel Embedding

For the kernel embedding, we again utilize the Gaussian kernel (30) with bandwidth $\sigma = 10^{-3}$, the choice of which will be justified later. Unlike for the Whitney embedding, the kernel embedding does not yield explicit representations of the embedded densities $\mu(p^{\tau}_{x_i})$, but instead by Algorithm 3.1 yields only the kernel distance matrix

$$D_{ij} = \left\| \mu(p^{t}_{x_i}) - \mu(p^{t}_{x_j}) \right\|_{\mathbb{H}},$$

which cannot be visualized directly. We thus apply the Multidimensional scaling (MDS) algorithm to D, in order to visualize the level of similarity between the embedded densities.

Given a distance matrix D, MDS generates points $z_i \in \mathbb{R}^k$ in a Euclidean space of a chosen dimension $k \in \mathbb{N}$ such that the pairwise distance between the z_i optimally corresponds to the distances in D. For an overview of different MDS methods, see for example (Young 2013). We here use the implementation of classical MDS given by the `cmdscale` method in MATLAB.

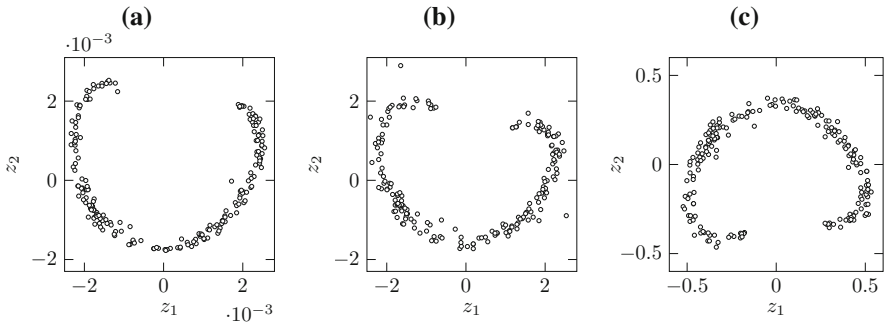

Fig. 6 MDS representations of different distance matrices between the transition densities. **a** Kernel distance matrix between embedded transition densities $\mu(p_x^t)$. The point cloud is a representations of the fuzzy transition manifold embedded into $\mathbb{H}$. **b** $L^2_{1/\rho}$ distance matrix between transition densities p_x^t. **c** L^2 distance matrix between transition densities p_x^t. Note that the MDS embedding is unique only up to distance preserving transformations, hence the difference in orientation here

The MDS representation of the kernel distance matrix for $k = 2$ is shown in Fig. 6a. The horseshoe structure of the MEP is immediately visible. Moreover, it is also possible to visualize the corresponding $L^2_{1/\rho}$ and L^2 distance matrices via MDS, i.e., the matrices

$$\left(D_{L^2_{1/\rho}}\right)_{ij} := \left\| p^t_{x_i} - p^t_{x_j} \right\|_{L^2_{1/\rho}} \quad \text{and} \quad \left(D_{L^2}\right)_{ij} := \left\| p^t_{x_i} - p^t_{x_j} \right\|_{L^2}.$$

The results are shown in Fig. 6b, c.

The MDS representation of D is structurally very similar to $D_{L^2_{1/\rho}}$ and D_{L^2}, up to scaling and rotation. This suggests that the $L^2_{1/\rho}$ and L^2 distances are preserved very well under μ, up to a constant factor. To confirm this, we now compute the empirical maximum distortion of the $L^2_{1/\rho}$ metric based on the given test points, i.e., $\mathscr{D}_N(\mu) := \mathscr{C}_N(\mu)\mathscr{E}_N(\mu)$ where

$$\mathscr{C}_N(\mu) := \max_{\substack{i,j=1,\dots,N \\ i\neq j}} \frac{\|p^\tau_{x_i} - p^\tau_{x_j}\|_{L^2_{1/\rho}}}{\|\mu(p^\tau_{x_i}) - \mu(p^\tau_{x_j})\|_{\mathbb{H}}}, \qquad \mathscr{E}_N(\mu) := \max_{\substack{i,j=1,\dots,N \\ i\neq j}} \frac{\|\mu(p^\tau_{x_i}) - \mu(p^\tau_{x_j})\|_{\mathbb{H}}}{\|p^\tau_{x_i} - p^\tau_{x_j}\|_{L^2_{1/\rho}}}.$$

For large enough N, we expect $\mathscr{D}_N(\mu)$ to be a good estimator for the true distortion $\mathscr{D}(\mu)$.

The blue graph in Fig. 7 shows the dependence of the empirical distortion on the kernel parameter σ. Here the minimum is $\mathscr{D}_N(\mu) \approx 6.4$ at $\sigma \approx 10^{-3}$. Interpreting $\mathscr{D}_N(\mu)$ as the condition number of the kernel-based embedding problem, the problem can be described as reasonably well-conditioned.

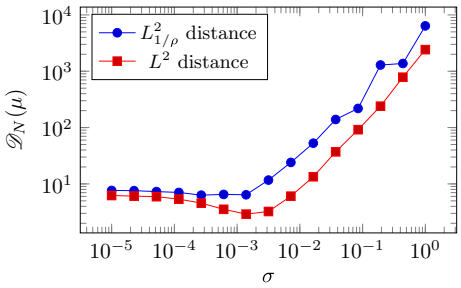

Fig. 7 Maximum distortion $\mathscr{D}(\mu)$ of the $L^2_{1/\rho}$ and L^2 distance under the kernel embedding μ for the Gaussian kernel depending on the kernel bandwidth σ

Analogously, we can define the empirical maximum distortion of the Whitney embedding as $\mathscr{D}_N(\mathcal{F}) := \mathscr{C}_N(\mathcal{F})\mathscr{E}_N(\mathcal{F})$, where

$$\mathscr{C}_N(\mathcal{F}) := \max_{\substack{i,j=1,\dots,N \\ i\neq j}} \frac{\|p^\tau_{x_i} - p^\tau_{x_j}\|_{L^2_{1/\rho}}}{\|\mathcal{F}(p^\tau_{x_i}) - \mathcal{F}(p^\tau_{x_j})\|_{\mathbb{R}^3}},$$

$$\mathscr{E}_N(\mathcal{F}) := \max_{\substack{i,j=1,\dots,N \\ i\neq j}} \frac{\|\mathcal{F}(p^\tau_{x_i}) - \mathcal{F}(p^\tau_{x_j})\|_{\mathbb{R}^3}}{\|p^\tau_{x_i} - p^\tau_{x_j}\|_{L^2_{1/\rho}}}.$$

For a given embedding $\mathcal{F}$, this distortion can again be computed numerically. For the "good" embedding $\mathcal{F}_g$, we obtain $\mathscr{D}_N(\mathcal{F}_g) \approx 5 \cdot 10^2$, while for the "bad" embedding $\mathcal{F}_b$, we obtain $\mathscr{D}_N(\mathcal{F}_b) \approx 7 \cdot 10^3$. The kernel embedding is therefore much better conditioned than both Whitney embeddings.

Remark 4.1 Analogously, we can also define and compute the maximum distortion $\mathscr{D}_N(\mu)$ of the L^2-metric (red graph in Fig. 7). Here, for $\sigma \approx 10^{-1}$ we obtain $\mathscr{D}_N(\mu) \approx 2.9$, i.e., the embedding becomes nearly isometric. This is not surprising as it has been shown in Sriperumbudur et al. (2010) that for radial kernels $k_\sigma(x, y) = \sigma^{-d} g(\sigma^{-1}\|x - y\|)$ where g is bounded, continuous, and positive definite, it holds that

$$\lim_{\sigma\to 0} \|\mu_{k_\sigma}(p) - \mu_{k_\sigma}(q)\|_{\mathbb{H}} = \|p - q\|_{L^2}.$$

The Gaussian kernel belongs to this class of kernels. We thus expect that by increasing the sample number M of the transition densities and further decreasing σ, the distortion can be reduced further. However, recall that for our application, only the distortion of the $L^2_{1/\rho}$ distance is relevant.

4.3 Alanine Dipeptide

We now demonstrate the applicability of Algorithm 3.1 to realistic, high-dimensional molecular systems by computing reaction coordinates of the Alanine dipeptide. The peptide, depicted in Fig. 8a, consists of 22 atoms, the state space $\mathbb{X}$ thus has the dimension $n = 66$.

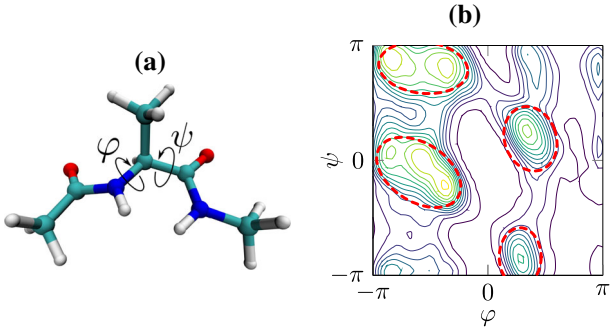

Fig. 8 The Alanine dipeptide. **a** Three-dimensional structure with the two essential dihedral angles (φ, ψ) highlighted. **b** The Ramachandran plot of (φ, ψ) reveals four local energy minima, i.e., metastable sets

We have chosen this molecule for our demonstrations as the mechanisms behind its long-term behavior, specifically its metastable sets and transition pathways, are well known, which helps to validate the results of our method. Moreover, the molecule is small enough for the relevant portions of state space to be sampled comprehensively, which is a requirement of our method. The Alanine dipeptide is also one of the most commonly used systems to demonstrate data-driven model reduction methods in molecular dynamics (Mardt et al. 2018).

The essential long-term behavior of this system is governed by the metastable transitions between four local minima of the potential energy surface (PES) (Chekmarev et al. 2004; Smith 1999). These minima are clearly visible when projecting the PES onto two specific backbone dihedral angles (φ, ψ) that we call *essential* from now on (see Fig. 8b). The transition between the metastable states happens along minimal energy pathways that we aim to reveal with our reaction coordinate. Note however that no information about the existence of the two essential dihedral angles was used in our experiments, and we perform all of the analysis on the full 66-dimensional data.

4.3.1 Setup and Parameter Choices

The simulations were performed using the Gromacs molecular dynamics software (Berendsen et al. 1995). We consider the molecule in explicit aqueous solution at temperature 400 K (the water molecules are discarded prior to further analysis). To generate the test points x_i, $N = 1000$ snapshots from a long, equilibrated trajectory were subsampled. This guarantees that the x_i cover the dynamically relevant regions of $\mathbb{X}$, i.e., the metastable sets and transition pathways. The values of the dihedral angles φ and ψ of the test points are shown in Fig. 10 (the x- and y-coordinates of the points). We see that the metastable sets and transition pathways from Fig. 8b are adequately covered. Note however that the projection onto the (φ, ψ)-space here serves only illustrative purposes; we continue to work with the test points in the full 66-dimensional space.

The intermediate lag time $\tau = 20$ ps falls between the slow and fast time scales of the system, which have been explicitly computed in Bittracher et al. (2018). In cases where a full time scale analysis is not available, the type and size of the molecule can often

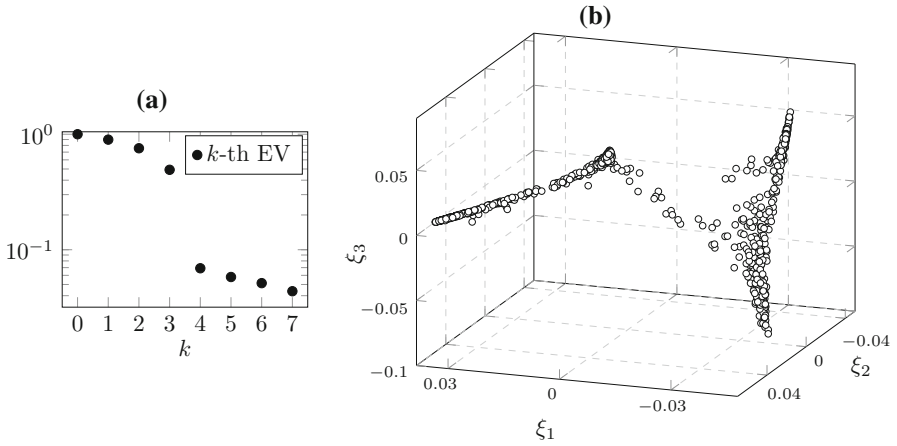

Fig. 9 Analysis of the kernel distance matrix D. **a** Eigenvalues of the diffusion map matrix. The existence of three eigenvalues close to 1 (not counting the eigenvalue 1 itself) indicates a three-dimensional reaction coordinate. **b** Test points in the space of the three sub-dominant diffusion map eigenvectors, i.e., the final three-dimensional reaction coordinate

hint at a suitable τ, as estimates for the time scales of certain common physiochemical reactions, such as dihedral angle reconfigurations or formation of secondary motifs, are well-known (Bittracher et al. 2018). For each test point x_i, $M = 96$ Gromacs MD simulations of length τ were performed, which took 40 h on a 96 core compute cluster. The resulting point clouds $\{y_i^{(l)}, l = 1, \ldots, M\}$ are samplings of the densities $p^{\tau}_{x_i}$.

To compute the kernel distance matrix D from the simulation data, the Gaussian kernel (30) with bandwidth $\sigma = 0.1$ was chosen. This particular value was chosen empirically to yield RCs that fit our expectations with regard to the transition pathways. However, the algorithm again appears to be very stable under variations of σ of even an order of magnitude. For the plug-in manifold learning algorithm that is applied to D, the diffusion maps algorithm with bandwidth $\sigma' = 0.01$ was used, again chosen empirically such that a clear low-dimensional manifold is visible in diffusion coordinante space (see Fig. 9b). For the choice of σ and σ', the same comments as in Sect. 4.1 apply.

The analysis of the simulation data was again performed in MATLAB and took 4 minutes on a standard 4 core laptop.

4.3.2 Results

Figure 9a shows the leading spectrum of the diffusion map matrix that was computed based on D. The first diffusion map eigenvalue is always equal to 1, and the associated eigenvector carries no structural information. Therefore, a spectral gap after the third sub-dominant eigenvalue indicates that the fuzzy transition manifold can be essentially parameterized by the three corresponding diffusion map coordinates in the sense of (1) and thus essentially be regarded as a three-dimensional object (provided that the eigenvectors are not higher-order modes, which is not the case here). The

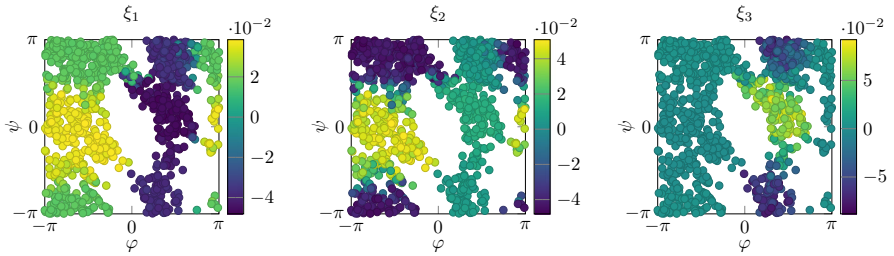

Fig. 10 The dihedral angles in the test points, colored by the three components of the reaction coordinate ξ. The coordinate ξ_1 primarily describes transitions in the angle φ, whereas ξ_2 and ξ_3 describe transitions in the angle ψ for low and high values of ψ, respectively

associated three subdominant eigenvectors now are the final reaction coordinate. For each of the 1000 test points, the values of the three eigenvectors are shown in Fig. 9b. This can be seen as the embedding of the test points into the reaction coordinate space. Here we observe four clusters of points, and three connecting paths. In Fig. 10, the values of the dihedral angles φ and ψ at the test points are compared to the values of the three components of the computed reaction coordinates, shown in color. We see that areas of almost constant color correspond to the four metastable sets from Fig. 8, and color gradients correspond to the transition pathways. Hence, the three-dimensional structure of the transition manifold corresponds to the network of transition pathways. Our reaction coordinate therefore accurately resolves transitions between metastable states.

5 Conclusion and Future Work

In this work, we have analyzed the embedding of manifolds that lie in certain function spaces into reproducing kernel Hilbert spaces. Moreover, we have proposed efficient numerical algorithms for learning parameterizations of these embedded manifolds from data. The question is motivated by the recent insight that parameterizations of the so-called transition manifold, a manifold consisting of the transition density functions of a stochastic system, are strongly linked to reduced coordinates for that system. The method can thus be used for coarse graining a given system.

Compared to previous approaches based on random embeddings into a Euclidean space, the new kernel-based approach eliminates the need to know the transition manifold dimension a priori. Furthermore, if a universal kernel is used, the topological structure of the transition manifold is guaranteed to be preserved under the embedding. We have derived bounds for the geometric distortion of the transition manifold under the RKHS embedding, which can be interpreted as the condition of the overall coarse graining procedure. Correspondingly, the numerical algorithm was demonstrated to be very robust, especially when compared to random embeddings, and, in realistic applications, we obtained very favorable results regarding algorithmic distortion bounds.

There are several new avenues to use the broader theory of kernel embeddings to characterize the kernel embedding of transition manifolds. First, we plan to improve the theoretic distortion bounds derived in Sect. 3.2 by considering different established interpretations of the metric defined by $d(p, q) = \|\mu(p) - \mu(q)\|_{\mathbb{H}}$. For an overview, see Sriperumbudur et al. (2010).

Recently, the spectral theory of transfer operators was extended to reproducing kernel Hilbert spaces in Klus et al. (2020). The usefulness of this new theory for the data-driven conformation analysis of molecular systems was demonstrated in Klus et al. (2018). As the transition manifold can be defined via the transfer operator[5], it seems natural to attempt to relate the embedded transition manifold to the kernel transfer operators and corresponding embedded transfer operators defined in Klus et al. (2020).

Finally, as illustrated in Bouvrie and Hamzi (2010), Bouvrie and Hamzi (2017a) and Bouvrie and Hamzi (2017b), RKHSs can act as *linearizing spaces* in the sense that performing linear analysis in the RKHS can capture strong nonlinearities in the original system. A typical example is the problem of linear separability in data classification: A data set which is not linearly separable might be easily separated when mapped into a nonlinear feature space. In our current context, this means that efficient linear manifold learning methods might be suitable to parameterize the embedded manifold, if the kernel is chosen appropriately. We will investigate whether a corresponding theory can be developed.

Acknowledgements The authors would like to thank the anonymous reviewers for constructive comments and suggestions that helped to improve the paper. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through Grant CRC 1114 "Scaling Cascades in Complex Systems", Project Number 235221301, Projects A01, B03, and B06.

Funding Open Access funding enabled and organized by Projekt DEAL.

A Proof of the Distortion Bounds

Proof of Proposition 3.11 First, note that $L^2_{1/\rho} \subset L^2$, as for $p \in L^2_{1/\rho}$

$$\|p\|_{L^2} = \|\sqrt{\rho}\, p\|_{L^2_{1/\rho}} \leq \|\sqrt{\rho}\|_\infty \|p\|_{L^2_{1/\rho}}. \tag{31}$$

[5] The fuzzy transition manifold is the image of all Dirac densities under the transfer operator, i.e., $\widetilde{\mathbb{M}} = \{\mathcal{T}^t \delta_x \mid x \in \mathbb{X}\}$.

Let (λ_i, φ_i) be the eigenpairs of the integral operator $\mathcal{T}_k$, ordered in decreasing order of λ_i. For arbitrary $p \in L^2_{1/\rho}$ consider the decomposition into the basis $\{\varphi_i\}_{i\in\mathbb{N}}$ of L^2:

$$p = \sum_{i=0}^{\infty} \tilde{p}_i \varphi_i.$$

Select $i_{\max} \in \mathbb{N}$ such that there exists an index $i \leq i_{\max}$ with $\tilde{h}_i \neq 0$, and such that $\lambda_i < \left(\varepsilon / \|\sqrt{\rho}\|_\infty\right)^2$ for all $i \geq i_{\max}$, and define

$$q = \sum_{i=0}^{i_{\max}-1} \tilde{p}_i \varphi_i.$$

Then,

$$\|p - q\|^2_{L^2} = \Big\| \sum_{i=i_{\max}}^{\infty} \tilde{p}_i \varphi_i \Big\|^2_{L^2} = \sum_{i=i_{\max}}^{\infty} \tilde{p}_i^2.$$

Further, using that $\{\sqrt{\lambda_i}\,\varphi_i\}_{i\in\mathbb{N}}$ forms an orthonormal basis of $\mathbb{H}$, and that μ is a linear operator, we get

$$\|\mu(p) - \mu(q)\|^2_{\mathbb{H}} = \Big\| \sum_{i=i_{\max}}^{\infty} \tilde{p}_i \lambda_i \varphi_i \Big\|^2_{\mathbb{H}} = \sum_{i=i_{\max}}^{\infty} \lambda_i \tilde{p}_i^2 \leq \lambda_{i_{\max}} \sum_{i=i_{\max}}^{\infty} \tilde{p}_i^2.$$

Thus we get

$$\frac{\|p - q\|_{L^2}}{\|\mu(p) - \mu(q)\|_{\mathbb{H}}} \geq 1/\sqrt{\lambda_{i_{\max}}} > \|\sqrt{\rho}\|_\infty / \varepsilon,$$

and with (31) finally

$$\frac{\|p - q\|_{L^2_{1/\rho}}}{\|\mu(p) - \mu(q)\|_{\mathbb{H}}} > \frac{1}{\varepsilon}.$$

□

Proof of Lemma 3.12 As $\{\sqrt{\lambda_i}\,\varphi_i\}_{i\in\mathbb{N}_0}$ forms an orthonormal basis of $\mathbb{H}$, we obtain

$$\|\mu(h)\|^2_{\mathbb{H}} = \Big\| \sum_{i=0}^{\infty} \tilde{h}_i \lambda_i \varphi_i \Big\|^2_{\mathbb{H}} = \sum_{i=0}^{\infty} \lambda_i \tilde{h}_i^2 \geq \sum_{i=0}^{i_{\max}} \lambda_i \tilde{h}_i^2.$$

Further, $\{\varphi_i\}_{i\in\mathbb{N}_0}$ forms an orthonormal basis of L_2, and so

$$\|h\|_2^2 = \Big\| \sum_{i=0}^{\infty} \tilde{h}_i \varphi_i \Big\|_2^2 = \sum_{i=0}^{\infty} \tilde{h}_i^2 = c(h, i_{\max}) \cdot \sum_{i=0}^{i_{\max}} \tilde{h}_i^2.$$

Thus,

$$\frac{\|\mu(h)\|_{\mathbb{H}}}{\|h\|_2} \geq \left(\frac{\sum_{i=0}^{i_{\max}} \lambda_i \tilde{h}_i^2}{c(h, i_{\max}) \cdot \sum_{i=0}^{i_{\max}} \tilde{h}_i^2} \right)^{1/2} \geq \sqrt{\frac{\lambda_{i_{\max}}}{c(h, i_{\max})}}.$$

□

Proof of Lemma 3.13 With assumption (24) and (25), we can write the left-hand side of (26) as

$$\begin{aligned} \|p_x^\tau\|_{L^2_{1/\rho}}^2 &= \int_{\mathbb{X}} p_x^\tau(y)^2 \frac{1}{\rho(y)} \mathrm{d}y \\ &< \int_{\mathbb{X}} \frac{\frac{1}{1-\delta(x,y)} \sum_{i=1}^d c_i(x) p_{A_i}^\tau(y)}{\sum_{i=1}^d b_i p_{A_i}^\tau(y)} |p_x^\tau(y)| \mathrm{d}y =: (\star). \end{aligned}$$

As for all $x \in \mathbb{X}$ it holds $c_i(x) \geq 0$ and $\sum_{i=1}^d c_i(x) \leq 1$, we obtain

$$\begin{aligned} \left\| \frac{\frac{1}{1-\delta(x,\cdot)} \sum_{i=1}^d c_i(x) p_{A_i}^\tau}{\sum_{i=1}^d b_i p_{A_i}^\tau} \right\|_\infty &\leq \frac{1}{1-\delta^*} \left\| \frac{\sum_{i=1}^d c_i(x) p_{A_i}^\tau}{\sum_{i=1}^d b_i p_{A_i}^\tau} \right\|_\infty \\ &\leq \frac{1}{1-\delta^*} \left\| \frac{\sum_{i=1}^d p_{A_i}^\tau}{\sum_{i=1}^d b_i p_{A_i}^\tau} \right\|_\infty \\ &\leq \frac{1}{(1-\delta^*) \min_i b_i}. \end{aligned} \tag{32}$$

With this, we can estimate the integral $(\star)$ as

$$(\star) \leq \int_{\mathbb{X}} \frac{1}{(1-\delta^*) \min_i b_i} |p_x^\tau(y)| \mathrm{d}y = \frac{1}{(1-\delta^*) \min_i b_i} \underbrace{\|p_x^\tau\|_{L^1}}_{=1}.$$

□

Proof of Lemma 3.14 The proof is completely analogous to the proof of Lemma 3.13, while in the estimate corresponding to (32) we use that

$$\max_i |c_i(x_1) - c_i(x_2)| \leq 1.$$

□

References

Abraham, I., Bartal, Y., Neiman, O.: Advances in metric embedding theory. Adv. Math. **228**(6), 3026–3126 (2011)

Baxter, J.R., Rosenthal, J.S.: Rates of convergence for everywhere-positive Markov chains. Stat. Probab. Lett. **22**(4), 333–338 (1995)

Berendsen, H., van der Spoel, D., van Drunen, R.: Gromacs: a message-passing parallel molecular dynamics implementation. Comput. Phys. Commun. **91**(1), 43–56 (1995)

Berry, T., Harlim, J.: Variable bandwidth diffusion kernels. Appl. Comput. Harmonic Anal. **40**(1), 68–96 (2016)

Best, R.B., Hummer, G.: Reaction coordinates and rates from transition paths. Proc. Natl. Acad. Sci. **102**(19), 6732–6737 (2005)

Bittracher, A., Koltai, P., Klus, S., Banisch, R., Dellnitz, M., Schütte, C.: Transition manifolds of complex metastable systems: theory and data-driven computation of effective dynamics. J. Nonlinear Sci. **28**(2), 471–512 (2017)

Bittracher, A., Banisch, R., Schütte, C.: Data-driven computation of molecular reaction coordinates. J. Chem. Phys. **149**(15), 154103 (2018)

Bouvrie, J., Hamzi, B.: Balanced reduction of nonlinear control systems in reproducing kernel Hilbert space. In: Proceedings of 48th Annual Allerton Conference on Communication, Control, and Computing, pp. 294–301 (2010)

Bouvrie, J., Hamzi, B.: Kernel methods for the approximation of some key quantities of nonlinear systems. J. Comput. Dyn. **4**(1), 1–19 (2017)

Bouvrie, J., Hamzi, B.: Kernel methods for the approximation of nonlinear systems. SIAM J. Control Optim. **55**(4), 2460–2492 (2017)

Bowman, G., Volez, V., Pande, V.S.: Taming the complexity of protein folding. Curr. Opin. Struct. Biol. **21**(1), 4–11 (2011)

Bowman, G.R., Pande, V.S., Noé, F. (eds.): An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Advances in Experimental Medicine and Biology, vol. 797. Springer, Berlin (2014)

Camacho, C.J., Thirumalai, D.: Kinetics and thermodynamics of folding in model proteins. Proc. Natl. Acad. Sci. **90**(13), 6369–6372 (1993)

Chekmarev, D.S., Ishida, T., Levy, R.M.: Long-time conformational transitions of alanine dipeptide in aqueous solution: continuous and discrete-state kinetic models. J. Phys. Chem. B **108**(50), 19487–19495 (2004)

Coifman, R.R., Kevrekidis, I.G., Lafon, S., Maggioni, M., Nadler, B.: Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. Multiscale Model. Simul. **7**(2), 842–864 (2008)

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., Zoubin, G.: Structure discovery in nonparametric regression through compositional kernel search. In: Dasgupta, S., McAllester, D.: (eds), Proceedings of the 30th International Conference on Machine Learning, Volume 28 of Proceedings of Machine Learning Research, pp. 1166–1174, Atlanta, Georgia, USA, 17–19 Jun (2013). PMLR

E, W., Vanden-Eijnden, E.: Towards a theory of transition paths. J. Stat. Phys. **123**(3), 503–523 (2006)

E, W., Ren, W., Vanden-Eijnden, E.: String method for the study of rare events. Phys. Rev. B **66**, 052301 (2002)

E, W., Ren, W., Vanden-Eijnden, E.: Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. J. Chem. Phys. **126**(16), 164103 (2007)

Elber, R., Bello-Rivas, J.M., Ma, P., Cardenas, A.E., Fathizadeh, A.: Calculating iso-committor surfaces as optimal reaction coordinates with milestoning. Entropy **19**(5), 219 (2017)

Freddolino, P.L., Harrison, C.B., Liu, Y., Schulten, K.: Challenges in protein folding simulations: timescale, representation, and analysis. Nat. Phys. **6**(10), 751 (2010)

Frewen, T.A., Hummer, G., Kevrekidis, I.G.: Exploration of effective potential landscapes using coarse reverse integration. J. Chem. Phys. **131**(13), 10B603 (2009)

Froyland, G., Gottwald, G.A., Hammerlindl, A.: A trajectory-free framework for analysing multiscale systems. Phys. D Nonlinear Phenom. **328**, 34–43 (2016)

Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. In: Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS'07, pp. 489–496 (2007)

Gaspar, P., Carbonell, J., Oliveira, J.L.: On the parameter optimization of support vector machines for binary classification. J. Integr. Bioinform. **9**(3), 33–43 (2012)
Gesùa, G.D., Lelièvre, T., Peutreca, D.L., Nectouxa, B.: Jump Markov models and transition state theory: the quasi-stationary distribution approach. Faraday Discuss. **195**, 469–495 (2016)
Gönen, M., Alpaydin, E.: Multiple kernel learning algorithms. J. Mach. Learn. Res. **12**(64), 2211–2268 (2011)
Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(Mar), 723–773 (2012)
Hunt, B., Kaloshin, V.: Regularity of embeddings of infinite-dimensional fractal sets into finite-dimensional spaces. Nonlinearity **12**(5), 1263–1275 (1999)
Klein, R.: Scale-dependent models for atmospheric flows. Annu. Rev. Fluid Mech. **42**(1), 249–274 (2010)
Klus, S., Bittracher, A., Schuster, I., Schütte, C.: A kernel-based approach to molecular conformation analysis. J. Chem. Phys. **149**(24), 244109 (2018)
Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., Noé, F.: Data-driven model reduction and transfer operator approximation. J. Nonlinear Sci. **28**, 985–1010 (2018)
Klus, S., Schuster, I., Muandet, K.: Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. J. Nonlinear Sci. **30**(1), 283–315 (2020)
Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**(1), 1–27 (1964)
Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: rank-based criteria. Neurocomputing **72**(7), 1431–1443 (2009)
Majda, A.J., Klein, R.: Systematic multiscale models for the tropics. J. Atmos. Sci. **60**(2), 393–408 (2003)
Mardt, A., Pasquali, L., Wu, H., Noé, F.: Vampnets for deep learning of molecular kinetics. Nat. Commun. **9**(1), 5 (2018)
Mattingly, J.C., Stuart, A.M.: Geometric ergodicity of some hypo-elliptic diffusions for particle motions. Markov Process. Relat. Fields **8**(2), 199–214 (2002)
McGibbon, R.T., Husic, B.E., Pande, V.S.: Identification of simple reaction coordinates from complex dynamics. J. Chem. Phys. **146**(4), 44109 (2017)
Melzer, T., Reiter, M., Bischof, H.: Nonlinear feature extraction using generalized canonical correlation analysis. In: Dorffner, G., Bischof, H., Hornik, K. (eds), Artificial Neural Networks—ICANN 2001, pp. 353–360 (2001)
Mercer, J.: Functions of positive and negative type, and their connection the theory of integral equations. Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci. **209**(441–458), 415–446 (1909)
Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B.: Kernel mean embedding of distributions: a review and beyond. Found. Trends Mach. Learn. **10**(1–2), 1–141 (2017)
Müller, K.: Reaction paths on multidimensional energy hypersurfaces. Angewandte Chemie Int. Ed. Engl. **19**(1), 1–13 (1980)
Munkres, J.R.: Topology, 2nd edn. Prentice Hall, Upper Saddle River (2000)
Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G.: Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Appl. Comput. Harmon. Anal. **21**(1), 113–127 (2006)
Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., Weikl, T.R.: Constructing the full ensemble of folding pathways from short off-equilibrium simulations. Proc. Natl. Acad. Sci. **106**(45), 19011–19016 (2009)
Owhadi, H., Yoo, G.R.: Kernel flows: from learning kernels from data into the abyss. J. Comput. Phys. **389**, 22–47 (2019)
Prinz, J.-H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J.D., Schütte, C., Noé, F.: Markov models of molecular kinetics: generation and validation. J. Chem. Phys. **134**(17), 174105 (2011)
Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
Schervish, M.J., Carlin, B.P.: On the convergence of successive substitution sampling. J. Comput. Graph. Stat. **1**(2), 111–127 (1992)
Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
Schölkopf, B., Smola, A., Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**(5), 1299–1319 (1998)
Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., Peters, J.: Computing functions of random variables via reproducing kernel Hilbert space representations. Stat. Comput. **25**(4), 755–766 (2015)

Schütte, C., Sarich, M.: Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches. Number 24 in Courant Lecture Notes. American Mathematical Society, Providence (2013)
Schwantes, C.R., Pande, V.S.: Modeling molecular kinetics with tICA and the kernel trick. J. Chem. Theory Comput. **11**(2), 600–608 (2015)
Smith, P.E.: The alanine dipeptide free energy surface in solution. J. Chem. Phys. **111**(12), 5568–5579 (1999)
Smola, A., Gretton, A., Song, L., Schölkopf, B.: A Hilbert space embedding for distributions. In: Proceedings of the 18th International Conference on Algorithmic Learning Theory, pp. 13–31. Springer-Verlag (2007)
Socci, N.D., Onuchic, J.N., Wolynes, P.G.: Diffusive dynamics of the reaction coordinate for protein folding funnels. J. Chem. Phys. **104**(15), 5860–5868 (1996)
Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.R.: Hilbert space embeddings and metrics on probability measures. J. Mach. Learn. Res. **11**, 1517–1561 (2010)
Steinwart, I., Christmann, A.: Support Vector Machines, 1st edn. Springer, New York (2008)
Vanden-Eijnden, E., Venturoli, M.: Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. J. Chem. Phys. **130**(19), 194103 (2009)
Young, F.W.: Multidimensional Scaling: History, Theory, and Applications. Psychology Press, New York (2013)
Zhang, W., Hartmann, C., Schütte, C.: Effective dynamics along given reaction coordinates, and reaction rate theory. Faraday Discuss. **195**, 365–394 (2016)
Zwanzig, R.: Memory effects in irreversible thermodynamics. Phys. Rev. **124**, 983–992 (1961)
Zwanzig, R.: Nonequilibrium Statistical Mechanics. Oxford University Press, Oxford (2001)