ESAIM: M2AN 43 (2009) 721–742
DOI: 10.1051/m2an/2009027

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

TOWARDS EFFECTIVE DYNAMICS IN COMPLEX SYSTEMS BY MARKOV KERNEL APPROXIMATION *

Christof Schütte[1] and Tobias Jahnke[2]

Abstract. Many complex systems occurring in various application share the property that the underlying Markov process remains in certain regions of the state space for long times, and that transitions between such metastable sets occur only rarely. Often the dynamics within each metastable set is of minor importance, but the transitions between these sets are crucial for the behavior and the understanding of the system. Since simulations of the original process are usually prohibitively expensive, the effective dynamics of the system, *i.e.* the switching between metastable sets, has to be approximated in a reliable way. This is usually done by computing the dominant eigenvectors and eigenvalues of the transfer operator associated to the Markov process. In many real applications, however, the matrix representing the spatially discretized transfer operator can be extremely large, such that approximating eigenvectors and eigenvalues is a computationally critical problem. In this article we present a novel method to determine the effective dynamics *via* the transfer operator *without* computing its dominant spectral elements. The main idea is that a time series of the process allows to approximate the sampling kernel of the process, which is an integral kernel closely related to the transition function of the transfer operator. Metastability is taken into account by representing the approximative sampling kernel by a linear combination of kernels each of which represents the process on one of the metastable sets. The effect of the approximation error on the dynamics of the system is discussed, and the potential of the new approach is illustrated by numerical examples.

Mathematics Subject Classification. 60J25, 60J35, 62M05, 60J22, 65C40.

Received July 7, 2008.
Published online July 8, 2009.

Introduction

This article deals with a novel approach to the identification of the effective dynamical behavior of complex systems. It will be assumed that the evolution of the system under consideration can be described by a Markov process. Furthermore we will be mainly interested in systems that (A) are high dimensional and (B) that exhibit metastable or almost invariant sets that are characterized by the property that the expected exit times

Keywords and phrases. Effective dynamics, complex systems, Markov process, metastability, transfer operators, model reduction, mixture models.

* *Supported by the DFG research center* Matheon *"Mathematics for key technologies" in Berlin and by Microsoft Research within the project "Conceptionalization of Molecular Dynamics".*

[1] Institut für Mathematik II, Freie Universität Berlin, Arnimallee 2–6, 14195 Berlin, Germany. schuette@math.fu-berlin.de

[2] Institut für Angewandte und Numerische Mathematik, Universität Karlsruhe (TH), Kaiserstr. 93, 76133 Karlsruhe, Germany. jahnke@math.uni-karlsruhe.de

Article published by EDP Sciences

from these sets define the timescales of the effective dynamical behavior of the system. Systems with these properties are pervasive; for example, they occur in the geo-sciences (*e.g.* climate dynamics with warm and ice ages or atmospheric blocking dynamics), in the economic sciences (*e.g.*, financial markets and their dynamical regimes or employment dynamics), or in the life sciences. Form the latter field our guiding example is taken: In biomolecular systems the metastable sets are called conformations and the effective dynamics can be described as transitions between these conformations with specific internal motion within each conformation [14,34]. Transitions between conformations are critical to the function of proteins and nucleic acids. They include ligand binding [26], complex conformational rearrangements between native protein substates [12,25], and folding [21], so that their analysis is the key to understanding the biophysics of living cells.

In many of these application the mathematical models used to describe the dynamics of the respective system are high dimensional ordinary differential equations (ODE). Mostly, appropriate models do not only contain the system's degrees of freedom but also additional degrees of freedom that represent the environment or heat bath in which the system is embedded or to which it is coupled. Considering the evolution of the system thus often means considering a process that results from the projection of some higher dimensional ODE to the system's state space. A typical example for this approach are thermostat models in molecular dynamics [23]. Under certain conditions (*e.g.*, on the time scales of the effective dynamics) the resulting process still is a Markov process. In many cases, however, the projected process is remodeled in form of a stochastic differential equation (SDE). Because of this, we will mainly consider SDE models for Markov processes on the state space of the system under consideration. Projected ODEs is briefly discussed in the appendix.

Mathematically each Markov process can be described by the associated Markov or *transfer operator*, and the effective dynamical behavior of the process by the essential properties (*e.g.* dominant eigenvalues and eigenvectors) of its transfer operator; there is a remarkably long list of articles about this topic from which the following may fit best to the topic considered herein [1,3,5,6,9,19,20,22,24,28,31,33]. Using the transfer operator in order to study the effective dynamics has one main advantage and one main disadvantage: the advantages is that independent of the (mostly strong) nonlinearity of the dynamics, the transfer operator is a *linear* operator that governs the propagation of a probability density by the underlying dynamics. The disadvantage, however, is that the transfer operator lives in a high dimensional function space whose dimension is that of the state space of the system. In many real-world cases, computation of its essential properties (dominant spectral elements) thus suffers from the curse of dimensionality although there are several articles that offer cures to this problem for specific problems mainly for biomolecular systems, see [4,29,30] for example.

This article provides a novel approach to the analysis of the effective dynamics *via* the transfer operator *without* computing its dominant spectral elements. Instead, we will consider the approximation of the integration kernel of the transfer operator. We will demonstrate that for non-deterministic dynamics this kernel has some nice structural properties that may allow to approximate it well even in high dimensional cases. We then will study how some mathematical properties of the dynamics, particularly its metastability properties, may change if we exchange the original transfer operator with the one that results from approximation of the kernel. Furthermore, we will introduce some algorithmic kernel approximation techniques that have the potential to work well even in high dimensions. Finally, we will present some numerical experiments to illustrate the concept itself and the performance of our kernel approximation techniques. However, it should be emphasized that this article can just give a first introduction of the key ideas: we will mainly consider diffusive dynamics in low dimensions for the sake of clarity and completeness, and we will base our kernel approximation algorithm on just one kind of ansatz functions. Generalizations are under investigation but will not be discussed herein.

1. Transfer operators and kernels

Throughout this article we study *homogeneous Markov processes* $X_t = \{X_t\}_{t\in\mathcal{I}}$ on a state space $\mathbf{X} \subset \mathbb{R}^d$, where $\mathcal{I}$ is an interval in $\mathbb{R}$. The dynamics of X_t is given by the stochastic transition function

$$p(t, x, A) = \mathbb{P}[X_{t+s} \in A \mid X_s = x], \tag{1.1}$$

for every $t, s \in \mathcal{I}$, $x \in \mathbf{X}$ and $A \subset \mathbf{X}$. We say that the process X_t admits an invariant probability measure μ on the corresponding measure space $(\mathbf{X}, \mathcal{A}, \mu)$, if

$$\int_{\mathbf{X}} p(t, x, A)\mu(\mathrm{d}x) = \mu(A) \quad \text{for all } A \in \mathcal{A}.$$

In the following we shall always assume that the invariant measure of the process exists and is unique. A Markov process is called reversible with respect to an invariant probability measure μ, if it satisfies

$$\int_A p(t, x, B)\mu(\mathrm{d}x) = \int_B p(t, x, A)\mu(\mathrm{d}x)$$

for every $t \in \mathcal{I}$ and $A, B \in \mathcal{A}$. If moreover $p(t, x, \cdot)$ is absolutely continuous with respect to the Lebesgue measure, then we denote by $\rho(t, x, y)$ the associated *flat-space transition density*, *i.e.*, we have

$$p(t, x, A) = \int_A \rho(t, x, y)\mathrm{d}y.$$

Transfer operator. Given some measure ν, we consider the function spaces

$$\begin{aligned} L^p_\nu &= \left\{ f : \mathbf{X} \to \mathbb{C} : \int |f(x)|^p \nu(\mathrm{d}x) < \infty \right\}, \\ L^p_\nu(\mathbf{X} \times \mathbf{X}) &= \left\{ f : \mathbf{X} \times \mathbf{X} \to \mathbb{C} : \int |f(x, y)|^p \nu(\mathrm{d}x)\nu(\mathrm{d}y) < \infty \right\} \end{aligned}$$

with $p = 1$ or $p = 2$. The associated norms will be denoted by $\| \cdot \|_{\nu,p}$. We will consider two cases: when ν stands for the Lebesgue measure we call the spaces *flat*, and when ν is equal to the invariant measure μ, then the spaces are called *weighted*. We define the *semigroup of Markov propagators* or forward transfer operators $P^t : L^r_\mu \to L^r_\mu$ with $t \in \mathcal{I}$ and $1 \leq r < \infty$ by

$$\int_A P^t f(y)\mu(\mathrm{d}y) \;=\; \int_{\mathbf{X}} f(x)p(t, x, A)\mu(\mathrm{d}x)$$

for any measurable $A \subset \mathbf{X}$. If μ is invariant under the dynamics X_t, then it is easy to see that the characteristic function $\mathbf{1}_{\mathbf{X}} \in L^1_\mu$ of the entire state space is an invariant density of P^t, *i.e.*, we have $P^t\mathbf{1}_{\mathbf{X}} = \mathbf{1}_{\mathbf{X}}$. As following from its definition, P^t conserves norm, $\|P^t f\|_1 = \|f\|_1$ and positivity, *i.e.*, $P^t f \geq 0$ whenever $f \geq 0$. Hence, P^t is a Markov operator. The perhaps simplest case is that of an ODE $\dot{z} = F(z)$. Let its solution be unique for all initial values $z(0) = z_0$ and denote its flow map by Φ^t such that $z(t) = \Phi^t z_0$. Then, the associated transfer function is $p(t, x, A) = \mathbf{1}_A(\Phi^t x)$ where $\mathbf{1}_A$ denotes the characteristic function of the set A. Let μ be some measure that is invariant under Φ^t. The corresponding transfer operator then reads $P^t f(x) = f(\Phi^{-t} x)$.

Basic assumption. In all of the subsequent we will suppose that $p(t, x, \cdot)$ as well as the associated invariant measure are absolutely continuous with respect to the Lebesgue measure. This simplifies our considerations. Needless to say that all of the subsequent definitions can be generalized to the case where absolute continuity cannot be assumed. Moreover, we tacitly assume that the invariant measure μ is nonzero almost everywhere. Again, this assumption can be dropped if the following arguments are restricted to the subset $\{x \in \mathbf{X}: \mu(x) > 0\} \subset \mathbf{X}$, but this would make the notation somewhat more complicated.

Kernels. With these assumption, the expression for the propagator P^t becomes

$$P^t f(y) \;=\; \int_{\mathbf{X}} k_t(y, x) f(x)\mu(x)\mathrm{d}x\,, \quad f \in L^p_\mu, \tag{1.2}$$

where $\mu(\mathrm{d}x) =: \mu(x)\mathrm{d}x$, and we have introduced the *transition kernel*

$$k_t(y,x)\mu(y) = \rho(t,x,y) \tag{1.3}$$

that is defined for all x, y for which $\mu > 0$. Obviously, the transition kernel satisfies

$$\int_{\mathbf{X}} k_t(y,x)\mu(y)\mathrm{d}y = 1\,, \quad \forall (x,t) \in \mathbf{X} \times \mathcal{I}\,. \tag{1.4}$$

A kernel with property (1.4) is called a *Markov kernel*. For a reversible process the transition kernel is symmetric, *i.e.*, $k_t(x,y) = k_t(y,x)$. We will furthermore consider a second kernel function, called the *sampling kernel*, being defined by

$$\kappa_t(x,y) = \mu(x)\,\rho(t,x,y) = \mu(x)k_t(y,x)\mu(y). \tag{1.5}$$

The sampling kernel is particularly important because it can be sampled directly from a given realization of the investigated process (see Sect. 2.4 for an example). In the following we will often fix a time t and then ignore the index t so that we simply can write the transition function as $\rho(x,y)$, the sampling kernel as $\kappa(x,y)$, and the transfer operator as $Pf(y) = \int_{\mathbf{X}} k(y,x)f(x)\mu(x)\mathrm{d}x$ in L^p_μ. For convenience we introduce the abbreviation $Pf = k * f$, knowing that we have to understand it relative to the space, especially weighting, considered.

2. Gaussian kernels and Ornstein-Uhlenbeck processes

2.1. **Ornstein-Uhlenbeck sampling kernels**

Consider an Ornstein-Uhlenbeck process

$$\dot{x} = -F(x - \bar{x}) + \Sigma\dot{W}, \tag{2.1}$$

with symmetric, positive definite matrices $F \in \mathbb{R}^{d\times d}$ and $\Sigma \in \mathbb{R}^{d\times d}$, and define $B = \Sigma^2$. The transition function of such a process is absolutely continuous with respect to the Lebesgue measure. The corresponding transition density at time t with respect to the initial condition to start in x_0 at time $t = 0$ is given by

$$\rho(t,x_0,x) = Z(t)\exp\left(-\frac{1}{2}(x - \xi(t))^T C(t)^{-1}(x - \xi(t))\right), \tag{2.2}$$

where

$$\begin{aligned} \xi(t) &= \bar{x} + \exp(-tF)(x_0 - \bar{x}) \\ Z(t) &= (2\pi)^{-d/2}(\det C(t))^{-1/2} \end{aligned}$$

and $C(t)$ is the solution of

$$C(t)F + FC(t) = B - \exp(-tF)B\exp(-tF). \tag{2.3}$$

It is well-known that for any $M_1, M_2, M_3 \in \mathbb{R}^{d\times d}$ the matrix equation $XM_1 + M_2X = M_3$ has a unique solution if $\lambda^{(1)} + \lambda^{(2)} \neq 0$ for all eigenvalues $\lambda^{(1)}$ of M_1 and $\lambda^{(2)}$ of M_2. Since F is positive definite, a unique solution of equation (2.3) exists. As a consequence of (2.2), the invariant measure is absolutely continuous with respect to the Lebesgue measure and has the form

$$\mu(x) = Z_\infty \exp\left(-\frac{1}{2}(x - \bar{x})^T C_\infty^{-1}(x - \bar{x})\right), \tag{2.4}$$

with C_∞ such that $C_\infty F + FC_\infty = B$. The last equation again has a unique solution; it satisfies

$$C(t) = C_\infty - \exp(-tF)C_\infty \exp(-tF). \tag{2.5}$$

The associated Markov operator P_t in the flat L^p space is obtained from $P_t f(x) = \int \rho(t, x_0, x) f(x_0) \mathrm{d}x_0$.

We consider the sampling kernel

$$\kappa_t(x_0, x) = \rho(t, x_0, x)\mu(x_0), \tag{2.6}$$

because this is the object that can be sampled directly from a given realization of the Ornstein-Uhlenbeck process (for details see Sect. 2.4 below). Equations (2.2) and (2.4) yield that the sampling kernel can be expressed as

$$\kappa_t(x_0, x) = Z(t)Z_\infty \exp\left(-\frac{1}{2}((x-\bar{x})^T, (x_0-\bar{x})^T)\mathcal{C}(t)^{-1}\begin{pmatrix} x-\bar{x} \\ x_0-\bar{x}\end{pmatrix}\right), \tag{2.7}$$

with

$$\mathcal{C}^{-1}(t) = \begin{pmatrix} C(t)^{-1} & -C(t)^{-1}\exp(-tF) \\ -\exp(-tF)C(t)^{-1} & \exp(-tF)C(t)^{-1}\exp(-tF) + C_\infty^{-1}\end{pmatrix}. \tag{2.8}$$

According to (1.5), the associated Markov kernel with respect to the invariant measure μ has the form

$$k_t(x, x_0) = \frac{1}{\mu(x)}\kappa_t(x_0, x)\frac{1}{\mu(x_0)},$$

and we observe that k_t is indeed symmetric (this follows from (2.5)) as it should be since the Ornstein-Uhlenbeck process is a reversible Markov process.

2.2. **Parameter estimation from a time series of the sampling kernel**

Suppose now that the parameters F, $\bar{x}$, and Σ of the process are unknown, and that only a sampling of the sampling kernel is given. Based on this sampling, the parameters of the Ornstein-Uhlenbeck process can be estimated as follows:

(1) Approximate the covariance matrix $\hat{\mathcal{C}} \approx \mathcal{C}$ of the sampling kernel and its inverse

$$\hat{\mathcal{C}}^{-1}(t) \;=\; \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^T & M_{22}\end{pmatrix}.$$

(2) Compute an estimate $\hat{F} \approx F$ by solving

$$\hat{F} \;=\; -\log(-M_{11}^{-1}M_{12})/t$$

where $\log(\cdot)$ denotes the matrix logarithm. Since $(-M_{11}^{-1}M_{12})$ is (at least approximately) a symmetric and positive definite matrix, the matrix logarithm is well-defined *via* the logarithm of the eigenvalues.

(3) Approximate $\hat{\Sigma} \approx \Sigma$ *via*

$$\begin{aligned} \hat{C}_\infty^{-1} &= M_{22} - M_{12}^T M_{11}^{-1} M_{12}, \\ \hat{B} &= \hat{C}_\infty \hat{F}^T + \hat{F}\hat{C}_\infty, \\ \hat{\Sigma} &= \sqrt{\hat{B}}. \end{aligned}$$

If $\hat{C}_\infty$ and $\hat{F}$ are symmetric and positive definite, then so is $\hat{B}$, and the matrix $\hat{\Sigma} = \sqrt{\hat{B}}$ can be obtained by computing the eigendecomposition of $\hat{B}$ and taking the square roots of the eigenvalues.

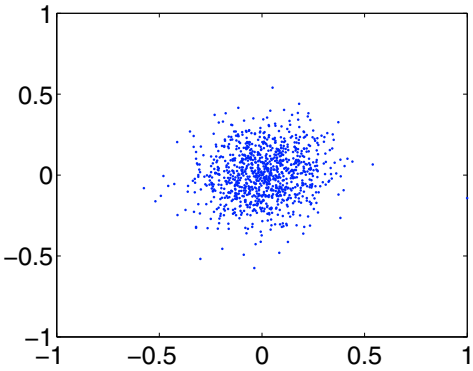

FIGURE 1. Sampling (as resulting from DNS) of the Ornstein-Uhlenbeck transition kernel for $F = 4$ and $\sigma = 0.45$ with $t = 0.5$ and $m = 1000$ steps.

2.3. **Invariant measure of sampling kernels**

Whenever a sampling kernel κ is known, then the associated invariant measure is computable by means of simple integration:

$$\mu(x) = \int \rho(x_0, x)\mu(x_0)\mathrm{d}x_0 = \int \kappa(x_0, x)\mathrm{d}x_0.$$

As in (1.5), $\rho(\cdot,\cdot)$ denotes the flat space transition function of the underlying process. For Gaussian sampling kernels

$$\kappa(x, x_0) \propto \exp\left(-\frac{1}{2}(x^T, x_0^T)\begin{pmatrix} M_{xx} & M_{xx_0} \\ M_{xx_0}^T & M_{x_0x_0} \end{pmatrix}\begin{pmatrix} x \\ x_0 \end{pmatrix}\right),$$

we thus get for the associated measure

$$\mu(x) \propto \exp\Big(-\frac{1}{2}x^T\mathcal{M}x\Big), \quad \text{with} \quad \mathcal{M} = M_{x_0x_0} - M_{xx_0}^T M_{xx}^{-1} M_{xx_0}. \tag{2.9}$$

The kernels are symmetric (and thus induce a reversible process) iff $M_{xx} = M_{x_0,x_0}$ and $M_{x_0x} = M_{x_0x}^T$.

2.4. **Numerical illustration**

Let us consider the 1d-case with $F = 4$, $\bar{x} = 0$ and $\Sigma = 0.45$. A direct numerical simulation (DNS) of the system (with Euler-Maruyama discretization in time with timestep 0.001) yields a time series $x_{0:m-1} = \{x_0, \ldots, x_m\}$ with $x_n \in \mathbf{X}$. If the DNS is *ergodic* in the sense that the generated ensemble of points in state space is approximately distributed according to the invariant measure μ of the process), then the sampling $x_{0:m-1}$ *directly* induces a sampling $z_{0:m-1} = \{z_0, \ldots, z_{m-1}\}$ of the sampling kernel by letting $z_n = (x_n, x_{n+1}) \in \mathbf{X} \times \mathbf{X}$. This sampling is illustrated in Figure 1 for $t = 0.5$ and $m = 1000$ sampling points.

In the one-dimensional case, one obtains $C_\infty = \Sigma^2/2F$, $C(t) = C_\infty(1 - \exp(-2tF))$, and

$$\mathcal{C}^{-1}(t) = C^{-1}(t)\begin{pmatrix} 1 & -\exp(-tF) \\ -\exp(-tF) & 1 \end{pmatrix}.$$

Hence, the covariance matrix of the sampling kernel is

$$\mathcal{C}(t) = \frac{\Sigma^2}{2F}\begin{pmatrix} 1 & \exp(-tF) \\ \exp(-tF) & 1 \end{pmatrix}.$$

TABLE 1. Dependence of the estimators on the sampling length m.

m	1000	2000	10000	∞
$\hat{F}$	4.530	4.050	4.010	4.000
$\hat{\Sigma}$	0.503	0.456	0.452	0.450

An estimate $\hat{\mathcal{C}} = (\hat{\mathcal{C}}_{ij})_{i,j}$ of this matrix is directly available from the given data. The approximations $\hat{F} \approx F$ and $\hat{\Sigma} \approx \Sigma$ can be computed *via*

$$\hat{F} = \log\left(\hat{\mathcal{C}}_{11}/\hat{\mathcal{C}}_{12}\right)/t, \qquad \hat{\Sigma}_N = \sqrt{2\hat{F}_N\hat{\mathcal{C}}_{11}}.$$

Table 1 shows that the estimates converge to the correct values as the length m of the time series increases.

The example shows that the sampling kernel allows to estimate the parameters of the underlying Ornstein-Uhlenbeck process in an easy way. However, in many realistic applications, the underlying SDE does not have the simple form (2.1), because the process is metastable. Nevertheless, it will be shown how such situations can be treated by a superposition of kernels each of which represents an Ornstein-Uhlenbeck process.

3. METASTABILITY

Let us make a simple Gedankenexperiment (experiment of thought): Assume we consider a diffusion process x_t in state space $\mathbb{R}$, governed by the SDE

$$\dot{x}_t = -F(x_t) + \Sigma\dot{W}_t,$$

where $F = -DV(x)$ is the gradient of a smooth potential V with several local minima. Then it is well known that for small enough Σ the process stays for long periods of time in the disjoint wells $M_1, \ldots, M_N \subset \mathbb{R}$ of V while the exit from the well M_i has an expected exit time that scales like $\exp(-2\Delta V_i/\Sigma^2)$, where ΔV_i is the lowest energy barrier from the respective well into a neighboring one. That is, the process is metastable and the sampling kernel of this process will then have the following approximate form: it will be a superposition of "peaks" in each of the wells, *i.e.*, in $M_i \times M_i$, $i = 1, \ldots, N$. Such processes occur in many real-life applications; examples have been given in the introduction. This is our motivation to study Markov processes that belong to sampling kernels constructed by superposition, and subsequently analyse their metastability properties.

3.1. **Superposition kernels**

Let us now discuss how to construct Markov operators and kernels by superposition. Suppose that P_i, $i = 1, \ldots, N$ are Markov operators in $L^p_{\mu_i}$ with respective invariant probability measures μ_i, and that $\alpha_i \in \mathbb{R}$ are non-negative weights with $\alpha_1 + \ldots + \alpha_N = 1$. In this case, we consider the Markov operator P in L^p_μ given by the superposition

$$\int_A Pf(x)\mu(\mathrm{d}x) = \sum_{i=1}^N \alpha_i \int_A P_i f(x)\mu_i(\mathrm{d}x), \quad \text{with} \quad P_i f(x) = \int k_i(x,y)f(y)\mu_i(\mathrm{d}y),$$

where μ is the invariant probability measure of P. These facts guarantee that $P_i\mathbf{1}_\mathbf{X} = \mathbf{1}_\mathbf{X} \in L^p_{\mu_i}$ and $P\mathbf{1}_\mathbf{X} = \mathbf{1}_\mathbf{X} \in L^p_\mu$. Inserting this into the above equation yields that μ is the invariant probability measure of P if and only if

$$\mu = \sum_{i=1}^N \alpha_i\mu_i. \tag{3.1}$$

The kernel associated with P is then given by

$$\mu(x)k(x,y)\mu(y) = \sum_{i=1}^{N} \alpha_i \mu_i(x) k_i(x,y) \mu_i(y), \tag{3.2}$$

where k_i denotes the Markov kernel of P_i, and it is assumed that all of the invariant measures are absolutely continuous with respect to the Lebesgue measure. A kernel with this structure will be called a superposition kernel.

The kernels are living in the respectively weighted spaces. What is the flat space transition density $\rho_{\text{flat}}(\cdot,\cdot)$ if the flat space transition densities $\rho_{i,\text{flat}}(\cdot,\cdot)$ belong to the kernels k_i? By using (1.5) and (3.2), we find the answer

$$\rho_{\text{flat}}(x,y) = \sum_{j=1}^{N} \alpha_j \frac{\mu_j(x)}{\mu(x)} \rho_{j,\text{flat}}(x,y), \tag{3.3}$$

where we assumed that μ is positive (almost) everywhere.

Remark 3.1. A realization of the process corresponding to the superposition kernel can be computed by repeating the following steps. Draw a random variable r from the uniform distribution $[0,1]$ and choose the index j such that $r \in [\beta_{j-1}(x), \beta_j(x))$, where $\beta_j(x) = \sum_{k=1}^{j} \alpha_k \mu_k(x)/\mu(x)$, and $j = N$ if $r = 1$. Then, the current state x is updated according to the jth transition density $\rho_{j,\text{flat}}(x,y)$. However, we emphasize that our goal is to solve the inverse problem: How can the parameters α_i and the densities ρ_i be estimated if a realization of the stochastic process is given? Before this question is addressed, we investigate the relation between superposition kernels and metastability.

3.2. Almost invariance and superposition kernels

As above let k_i, $i = 1,\dots,N$, be Markov kernels with invariant probability measures μ_i. We consider the mixed kernel $\mu(x)k(x,y)\mu(y) = \sum_{i=1}^{N} \alpha_i \mu_i(x) k_i(x,y) \mu_i(y)$, with the invariant measure $\mu = \sum_i \alpha_i \mu_i$ and weights α_i such that $\sum_i \alpha_i = 1$. Let us assume that the measures μ_i all are absolutely continuous with respect to the Lebesgue measure, and that μ is positive (almost) everywhere.

Definition 3.2. The Markov kernel k is called ϵ-metastable if

$$O_{ij} = \int \frac{\mu_i(x)\mu_j(x)}{\mu(x)} \,\mathrm{d}x \le \epsilon$$

for all $i,j = 1,\dots,N$ with $j \neq i$.

Almost invariant densities of k. In the μ-weighted space, the invariant densities of the Markov operators P_i are

$$\Phi_i(x) = \mu_i(x)/\mu(x).$$

Obviously, these functions are densities in L^1_μ. A nice property of these densities is formulated in the following lemma.

Lemma 3.3. *The kernel k is ϵ-metastable if and only if*

$$O_{ij} = \int \Phi_i(x)\Phi_j(x)\mu(x)\,\mathrm{d}x = \langle \Phi_i, \Phi_j \rangle_\mu \le \epsilon \tag{3.4}$$

for all $i,j = 1,\dots,N$ with $j \neq i$.

The proof follows directly from Definition 3.2.

Other useful properties are:

$$\begin{aligned}
\sum_{j=1}^{N} \alpha_j \Phi_j(x) &= 1, \quad \forall x, \ \textit{i.e.}, \ \{\Phi_j\}_{j=1,\dots,N} \text{ is a partition of unity,} \\
0 \le \Phi_i(x) &\le 1/\alpha_i, \quad \forall i, \qquad (3.5) \\
\Big|\Phi_i(x) - \frac{1}{\alpha_i}\Big|\mu(x) &= \frac{1}{\alpha_i}\Big(\sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_j \mu_j(x)\Big). \qquad (3.6)
\end{aligned}$$

The latter equation follows from

$$\Big|\Phi_i(x) - \frac{1}{\alpha_i}\Big|\mu(x) \;=\; \Big|\mu_i(x) - \sum_{j=1}^{N} \frac{\alpha_j}{\alpha_i}\mu_j(x)\Big| \;=\; \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{\alpha_j}{\alpha_i}\mu_j(x).$$

With these properties we can prove that the propagator with kernel k leaves the density Φ_i nearly invariant with respect to the measure μ:

Theorem 3.4. *Let the kernel k defined as above be ϵ-metastable, let all measures μ_i be absolutely continuous with respect to the Lebesgue measure, and let μ be positive (almost) everywhere. Then, for all $i = 1, \dots, N$:*

$$\|k * \Phi_i - \Phi_i\|_{1,\mu} \le 2(1 - \alpha_i)\,\epsilon$$

where $(k * \Phi_i)(y) = \int_{\mathbf{X}} k(y,x)\Phi_i(x)\mu(x)\mathrm{d}x$.

Proof. The proof requires some preparations:

$$\begin{aligned}
\|k * \Phi_i - \Phi_i\|_{1,\mu} &= \int \Big| \int \mu(y)k(y,x)\Phi_i(x)\mu(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\Big|\,\mathrm{d}y \\
&= \int \Big| \sum_j \alpha_j \int \mu_j(y)k_j(y,x)\Phi_i(x)\mu_j(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\Big|\,\mathrm{d}y \\
&= \int \Big| \sum_j \alpha_j\mu_j(y) \int k_j(y,x)\Phi_j(x)\mu_i(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\Big|\,\mathrm{d}y \\
&\le \int \sum_{j=1,j\neq i}^{N} \alpha_j \int k_j(y,x)\Phi_j(x)\mu_i(x)\mathrm{d}x\,\mu_j(y)\mathrm{d}y \\
&\quad + \int \Big|\alpha_i\mu_i(y) \int k_i(y,x)\Phi_i(x)\mu_i(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\Big|\,\mathrm{d}y.
\end{aligned}$$

The first of these terms can be estimated using that $\int k_j(y,x)\mu_j(y)\mathrm{d}y = 1$:

$$\int \sum_{j=1,j\neq i}^{N} \alpha_j \int k_j(y,x)\Phi_j(x)\mu_i(x)\mathrm{d}x\,\mu_j(y)\mathrm{d}y \le \sum_{j=1,j\neq i}^{N} \alpha_j \langle \Phi_i, \Phi_j\rangle_\mu \le (1-\alpha_i)\epsilon.$$

The second term is split into two parts:

$$\begin{aligned}\int \left|\alpha_i\mu_i(y)\int k_i(y,x)\Phi_i(x)\mu_i(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\right|\mathrm{d}y \;\;\leq\;\; & \int \left|\alpha_i\mu_i(y)\int k_i(y,x)\frac{1}{\alpha_i}\mu_i(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\right|\mathrm{d}y \\ & + \alpha_i\int\int k_i(y,x)\left|\Phi_i(x)-\frac{1}{\alpha_i}\right|\mu_i(x)\mathrm{d}x\,\mu_i(y)\mathrm{d}y.\end{aligned}$$

The first part vanishes because

$$\begin{aligned}\int \left|\mu_i(y)\int k_i(y,x)\mu_i(x)\mathrm{d}x \;-\; \Phi_i(y)\mu(y)\right|\mathrm{d}y \;\;&=\;\; \int\left|\int k_i(y,x)\mu_i(x)\mathrm{d}x \;-\; \mathbf{1}_{\mathbf{X}}\right|\mu_i(y)\,\mathrm{d}y \\ &=\;\; \int\left|P_i1(y) \;-\; \mathbf{1}_{\mathbf{X}}\right|\mu_i(y)\,\mathrm{d}y = 0\end{aligned}$$

since $P_i\mathbf{1}_{\mathbf{X}} = \mathbf{1}_{\mathbf{X}}$ in $L^2_{\mu_i}$. The second part allows the following estimate based on (3.6) and the fact that $\int k_i(y,x)\mu_i(y)\mathrm{d}y = 1$:

$$\begin{aligned}\int\int k_i(y,x)\left|\Phi_i(x)-\frac{1}{\alpha_i}\right|\mu_i(x)\mathrm{d}x\mu_i(y)\mathrm{d}y \;\;&\leq\;\; \int\frac{1}{\alpha_i}\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_j\mu_j(x)\right)\Phi_i(x)\mathrm{d}x \\ &\leq\;\; \frac{1}{\alpha_i}\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_j\int\mu_j(x)\Phi_i(x)\mathrm{d}x \\ &\leq\;\; \frac{1}{\alpha_i}(1-\alpha_i)\epsilon.\end{aligned}$$

Putting the two terms together again we get

$$\|k*\Phi_i-\Phi_i\|_{1,\mu}\leq(1-\alpha_i)\epsilon+(1-\alpha_i)\epsilon. \qquad \square$$

Let us now assume that K is a Markov kernel which cannot be represented *exactly* by a weighted sum of other kernels k_i but which can be *approximated* by such a representation, *i.e.*

$$\mu(x)K(x,y)\mu(y)\approx\mu(x)k(x,y)\mu(y)=\sum_{i=1}^{N}\alpha_i\mu_i(x)k_i(x,y)\mu_i(y).$$

As before, let μ_i be the invariant measures of the Markov kernels k_i and let μ be the invariant measure of k. It is assumed that μ is positive almost everywhere, and that all measures are absolutely continuous with respect to the Lebesgue measure. The coefficients α_i are positive with $\sum_j\alpha_j=1$.

Definition 3.5. The kernel K is said to be (ϵ,δ)-metastable if there is an ϵ-metastable kernel k that satisfies the above assumptions such that

$$\|K-k\|_{1,\mu}=\int\int|K-k|(y,x)\mu(x)\mathrm{d}x\mu(y)\mathrm{d}y\;\leq\delta.$$

Now we again consider the functions $\Phi_i(x) = \mu_i(x)/\mu(x)$ as candidates for almost invariant densities. By definition, we obtain

$$\begin{aligned}\|K * \Phi_i - \Phi_i\|_{1,\mu} &\leq \|k * \Phi_i - \Phi_i\|_{1,\mu} \\ &\quad + \int\int |K-k|(y,x)\Phi_i(x)\mu(x)\mathrm{d}x\,\mu(y)\mathrm{d}y,\end{aligned}$$

where the first term can be estimated due to Theorem 3.4, while the second term can be simplified again if K is (ϵ,δ)-metastable:

$$\int\int |K-k|(y,x)\Phi_i(x)\mu(x)\mathrm{d}x\,\mu(y)\mathrm{d}y \leq \frac{1}{\alpha_i}\|K-k\|.$$

That is, we have shown the following result:

Theorem 3.6. *Let K be an (ϵ,δ)-metastable Markov kernel with associated mixed kernel*

$$\mu(y)k(x,y)\mu(x) = \sum_{j=1}^{N} \alpha_j \mu_j(y) k_j(x,y) \mu_j(x).$$

Then the bounded functions $\Phi_i(x) = \mu_i(x)/\mu(x)$ are almost invariant under K:

$$\|K * \Phi_i - \Phi_i\|_{1,\mu} \leq 2(1-\alpha_i)\,\epsilon + \frac{\delta}{\alpha_i}.$$

It should be pointed out that in Theorems 3.4 and 3.6 the invariance is measured with respect to the weighted norm $\|\cdot\|_{1,\mu}$. This has important consequences which will be discussed below (*cf.* Sect. 5.1).

3.3. **Perturbation of the spectrum**

Which consequence does the approximation of the kernel function have for the eigenvalues of the associated operators? This question can be answered, at least under additional assumptions. Assume that the original transfer operator P has invariant measure μ and kernel $K(\cdot,\cdot)$, and that the associated Markov process is reversible. Then, P is self-adjoint in the Hilbert space L^2_μ. Furthermore, consider a symmetric approximation $k(\cdot,\cdot) \approx K(\cdot,\cdot)$. Let k induce the operator $\tilde{P}$ with the same invariant measure μ. This assumption simplifies the analysis, because now both P and $\tilde{P}$ can be considered as operators in L^2_μ. Next, we assume that $P - \tilde{P}$ is a Hilbert-Schmidt operator which is the case if and only if

$$\|K-k\|^2_{2,\mu} = \int\int |K-k|^2(y,x)\mu(x)\mathrm{d}x\mu(y)\mathrm{d}y \leq \infty. \tag{3.7}$$

In this situation, Theorem 3 of [11] applies and yields:

Corollary 3.7. *The above assumptions on P, $\tilde{P}$, K and k imply that there exist enumerations $\{\lambda_i\}$, and $\{\nu_i\}$ of the spectra of P and $\tilde{P}$, respectively, in L^2_μ, such that*

$$\sum_{i=1}^{\infty} |\lambda_i - \nu_i|^2 \leq \|K-k\|^2_{2,\mu}.$$

The corollary indicates that if K and its approximation k are close enough in Hilbert-Schmidt norm then the spectra of the associated transfer operators are very similar. Since the dominant eigenvalues define the most important metastable timescales, this means that the approximation of the kernel of a transfer operator at least has the potential of approximating the effective dynamics also. However, the result does not imply

that the Hilbert-Schmidt norm is the appropriate norm for getting optimal approximation of the dominant part of the spectrum.

Remark 3.8. The above results on the approximation of essential features of the dynamics lead to two different norms, $\|\cdot\|_{1,\mu}$ and $\|\cdot\|_{2,\mu}$. Let k be a kernel and κ the associated sampling kernel. Then, these norms read

$$\begin{aligned} \|k\|_{1,\mu} &= \int\int |\kappa(x,y)| \mathrm{d}x\,\mathrm{d}y \\ \|k\|_{2,\mu} &= \left(\int\int \left(\frac{\kappa(x,y)}{\mu(x)^{1/2}\mu(y)^{1/2}} \right)^2 \mathrm{d}x\,\mathrm{d}y \right)^2, \end{aligned}$$

that is, 1-norm approximation means sampling kernel approximation in an unweighted (Lebesgue) measures sense, while 2-norm means sampling kernel approximation with a weighting which is large where the invariant measure is small.

The results of this section mean, roughly speaking, that an (ϵ,δ)-metastable kernel K can be approximated by a superposition of kernels $k_1,\ldots,k_N$, and that the corresponding invariant measures $\mu_1,\ldots,\mu_N$ allow to construct functions $\Phi_1,\ldots,\Phi_N$ which are almost invariant densities of K in the weighted space. It will now be our goal to construct such an approximative kernel in a situation where K is not known explicitly and only a sampling of the associated process is given. We will see that under appropriate conditions this process can *locally* be approximated by the linear Ornstein-Uhlenbeck process discussed in Section 2. The goal is then to find the parameters (means and covariance matrices) of these local Ornstein-Uhlenbeck processes, and to combine these processes in such a way that the metastability behaviour is correctly reproduced.

4. Kernel approximation by mixture models

Let us assume that we have a sampling $z_{0:m-1} = \{z_0,\ldots,z_{m-1}\}$ of the sampling kernel κ of the Markov process X_t under consideration ($z_n \in \mathbf{X}\times\mathbf{X}$ for all $n = 0,\ldots,m-1$). At the moment it is of no importance whether this sampling results from a long-term observation of X_t or from an ensemble of short-term observations. We want to exploit the sampling $z_{0:m-1}$ in order to get an approximation of κ by a superposition kernel relative to the Lebesgue measure (*i.e.* in the 1-norm sense with respect to the associated kernel function, see Rem. 3.8).

To this end we assume that $z_{0:m-1}$ is an observation of m independent and identically distributed realizations of some random variable Z that is distributed according to a *mixture model*, *i.e.*, the density $\rho(z|\theta)$ of the probability $\mathbb{P}(Z\in A|\theta) = \int_A \rho(z|\theta)\mathrm{d}z$ for measurable sets $A\in\mathbf{X}\times\mathbf{X}$ has the form of a weighted sum of N component densities:

$$\rho(z|\theta) = \sum_{y=1}^{N} \rho(z|y,\theta)\mathbb{P}(Y=y|\theta).$$

Here, θ denotes unknown parameters that determine the form of the probability densities. For example, letting the component densities $\rho(Z|y,\theta)$, $y=1,\ldots,N$ be Gaussian, then $\rho(Z|\theta)$ is a weighted sum of N Gaussian, and we speak of a *Gaussian mixture model*. Note that we treat Y as a *random variable* whose value $y\in 1,\ldots,N$ we do not know. Therefore, Y is called the *hidden* assignment variable for Z to one of the component densities, and $\mathbb{P}(Y=y|\theta)$ is the probability of a certain value y of Y given θ. If we knew the value of Y, say $Y=y$, then the density for Z was just the y-th component density.

According to our assumption, each of our independent samples z_n of Z has its own assignment value y_n, *i.e.*, the realization z_n of Z comes from the realization y_n of Y. However, by this assumption, for given θ, the $y_0,\ldots,y_{m-1}$ are statistically independent, as well as the $z_0,\ldots,z_{m-1}$. Our aim is to choose the parameters θ such that the likelihood

$$\mathcal{L}(\theta \mid z_{0:m-1}) = \prod_{n=0}^{m-1} \rho(z_n \mid \theta) = \prod_{n=0}^{m-1}\sum_{y=1}^{N} \rho(z_n|y,\theta)\mathbb{P}(Y_{t_n}=y|\theta)$$

is maximized, *i.e.*, we seek the parameters for which the probability of the given observation over the family of models under consideration is maximal. Thus, the maximum likelihood estimate (MLE) $\hat{\theta}$ satisfies

$$\hat{\theta} = \operatorname{argmax}_\theta \mathcal{L}(\theta \mid z_{0:m-1}).$$

Equivalently, the MLE can be defined *via* the logarithm of $\mathcal{L}$, which according to the assumed statistical independencies, reads

$$\begin{aligned} \hat{\theta} &= \operatorname{argmax}_\theta \log \mathcal{L}(\theta \mid z_{0:m-1}) \\ \text{with } \log \mathcal{L}(\theta \mid z_{0:m-1}) &= \sum_{n=0}^{m-1} \log \left(\sum_{y=1}^{N} \rho(z_n|y,\theta) \mathbb{P}(Y_{t_n} = y|\theta) \right). \end{aligned} \tag{4.1}$$

Here, however, we do not have observations of Y_t, *i.e.*, in some sense we have to consider the optimization task for all possible probability densities for Y_t. This is indeed feasible by means of the *expectation-maximization* (EM) algorithm. Hartley [15] pioneered the research on the EM algorithm in the late 1950s, followed by Dempster *et al.* [7] in the late 1970s. Over the years, the EM algorithm has found many applications in various domains and has become a powerful estimation tool [10,13].

The EM algorithm is an iterative optimization procedure. Starting with an initial parameter estimate $\theta^{(0)}$, each iteration monotonically increases the likelihood function $\mathcal{L}(\theta|x) = \mathcal{L}(\theta)$. Each iteration consists of two steps: (a) the E-step or the *expectation* step and (b) the M-step or the *maximization* step.

EM for the Gaussian mixture model. Let us now specify that the component densities are Gaussian with mean $\bar{z}_y$ and covariance matrices Σ_y, $y = 1, \ldots, N$:

$$\rho(z\,|\,y,\theta) = G(z; \bar{z}_y, \Sigma_y). \tag{4.2}$$

Thus, the free parameters θ of our model are the means and covariances of the Gaussian component densities, and the probabilities that component y is active,

$$\alpha_y = \mathbb{P}(Y = y\,|\,\theta). \tag{4.3}$$

These probabilities obviously have to satisfy the constraint $\sum_{y=1}^{N} \alpha_y = 1$. We thus have the parameter set

$$\theta = (\bar{z}_1, \ldots, \bar{z}_N, \Sigma_1, \ldots, \Sigma_N, \alpha_1, \ldots, \alpha_N).$$

In this case, the EM iteration takes the form of Algorithm 1.

The meaning of the algorithm becomes clearer with the interpretation of $\gamma_n^{(i)}(y)$ as the probability that according to the mixture model with parameters $\theta^{(i)}$ at time t_n the observation z_n has to be assigned to hidden state y. Thus, some sample z_n can be assigned to several of the hidden states with probability between 0 and 1.

Remark 4.1. Let $x_{0:m-1} = \{x_0, \ldots, x_m\}$ be some given observation sequence of the Markov process X_t under consideration. It induces a sampling $z_n = (x_n, x_{n+1}) \in \mathbf{X} \times \mathbf{X}$ of the associated sampling kernel κ. In this case the basic assumption of the mixture model that $z_{0:m-1}$ results from repeated independent and identically distributed realization of some random variable Y seems strange since we know that the z_n are correlated by the Markov process. Nevertheless the EM algorithm for the Gaussian mixture model often results in excellent approximations of the sampling (provided the underlying structure of the data is that of an superposition Gaussian kernel). However, one can take the Markov-like correlations in $z_{0:m-1}$ into account by means of generalizing the mixture model to *hidden Markov models* (HMM) which again leads to a specific version of the EM algorithm [2,27].

Algorithm 1 EM-algorithm for Gaussian mixture model

Require: Time series $z_{0:m-1} = \{z_0, \ldots, z_{m-1}\}$, tolerance tol, initial guess of parameters

$$\theta^{(0)} = (\bar{z}_y^{(0)}, \Sigma_y^{(0)}, \alpha_y^{(0)})_{y=1,\ldots,N}.$$

Ensure: Maximum likelihood estimate $\hat{\theta}$.

(1) Formally set $i = -1$.

(2) $i := i + 1$.

(3) Expectation step (E-step): Compute the occupation probabilities

$$\gamma_n^{(i)}(y) = \frac{G(z_n; \bar{z}_y^{(i)}, \Sigma_y^{(i)})\alpha_y^{(i)}}{\sum_{y=1}^N G(z_n; \bar{z}_y^{(i)}, \Sigma_y^{(i)})\alpha_y^{(i)}}, \quad y = 1, \ldots, N, \quad n = 0, \ldots, m-1$$

(4) Maximization step (M-Step):

For $y = 1, \ldots, N$, compute the new optimal parameter estimates

$$\begin{aligned}
\Sigma_y^{(i+1)} &= \frac{1}{\gamma^{(i)}(y)} \sum_{n=0}^{m-1} \gamma_n^{(i)}(y)(z_n - \bar{z}_y^{(i)})(z_n - \bar{z}_y^{(i)})^\top \\
\bar{z}_y^{(i+1)} &= \frac{1}{\gamma^{(i)}(y)} \sum_{n=0}^{m-1} \gamma_n^{(i)}(y)\, z_n \\
\alpha_y^{(i+1)} &= \frac{1}{\sum_{y=1}^N \gamma^{(i)}(y)} \gamma^{(i)}(y),
\end{aligned}$$

where $\gamma^{(i)}(y) = \sum_{n=0}^{m-1} \gamma_n^{(i)}(y)$.

(5) Compute the log-likelihood $\log cL(\theta^{(i+1)})$ using equations (4.1), (4.2), and (4.3).

If $\mathcal{L}(\theta^{(i+1)}) - \mathcal{L}(\theta^{(i)}) >$ tol, go to Step (2). Otherwise, terminate with $\theta^{(i+1)}$ as the desired approximation of $\hat{\theta}$.

Remark 4.2. Whenever we want to estimate the Gaussian parameters of the sampling kernel of some reversible Markov process it can be of interest to add appropriate symmetry constraints to the EM iteration. This is indeed possible by different means. The easiest way is, when considering a sampling $z_{0:m-1}$ with $z_n = (x_n, x_{n+1})$ that is induced by a long-term time series, to extend it into a reversible one by adding $z_{m:2m-1} = \{z_m, \ldots, z_{2m-1}\}$, $z_{n+m} = (x_n, x_{n+1})$, and then apply Algorithm 1 to the extended sampling.

5. Kernel approximation of metastable processes

We will now study metastable dynamics associated with the nonsymmetric double-well potential

$$V(x) = (x^2 - 1)^2 + 0.25x,$$

which is illustrated in Figure 2.

5.1. Diffusion in a double well potential

Let us return to a 1d Ornstein-Uhlenbeck process, this time of the form

$$\dot{x} = -F(x) + \Sigma \dot{W}$$

with $\Sigma = 0.45$, and force field $F(x) = \frac{\mathrm{d}V}{\mathrm{d}x}(x)$ where V denotes the above nonsymmetric double-well potential.

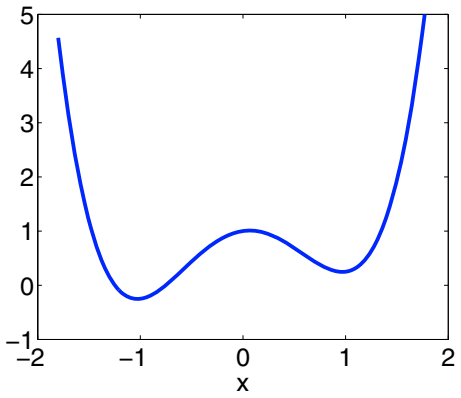

FIGURE 2. Nonsymmetric double well potential as used below for numerical tests.

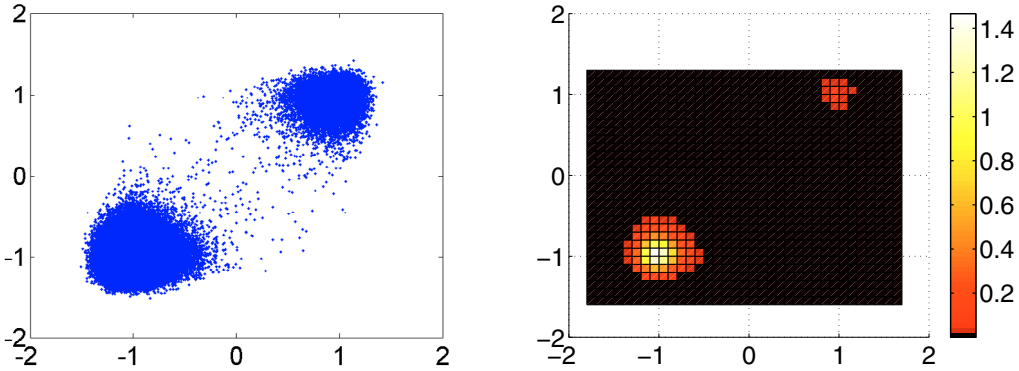

FIGURE 3. Sampling and histogram of sampling kernel for the OU process as discussed in the text. Points outside of the black box on the left hand side have not been hit by the sampling.

After again performing a DNS of the system (now with Euler-Maruyama discretization in time with timestep 0.001), we *directly* get a sampling of the sampling kernel $\kappa_t(x, x_0)$ with respect to the invariant measure μ of the process. For $t = 0.5$ and $N = 10\,000$ sampling points, the sampling is illustrated in the left panel of Figure 3. The right panel of Figure 3 shows the corresponding histogram.

We assume that the underlying kernel is (ϵ, δ)-metastable and apply Algorithm 1 to this sampling. This yields the following results for the mean values (estimated equilibria $(\bar{x}, \bar{x}_0)_i$ for $i = 1, 2$) and the respective covariance matrices $\Sigma_i(t)$, $i = 1, 2$:

$$(\bar{x}, \bar{x}_0)_1 = (-1.012, -1.012), \qquad (\bar{x}, \bar{x}_0)_2 = (0.933, 0.933),$$

and

$$\Sigma_1(t) = \begin{pmatrix} 0.0208 & 0.0010 \\ 0.0010 & 0.0211 \end{pmatrix}, \qquad \Sigma_2(t) = \begin{pmatrix} 0.0299 & 0.0047 \\ 0.0047 & 0.0308 \end{pmatrix}.$$

We observe that the two parts of the kernel can be well approximated by Gaussian that both have the structure and form of Gaussian kernels resulting from an Ornstein-Uhlenbeck process.

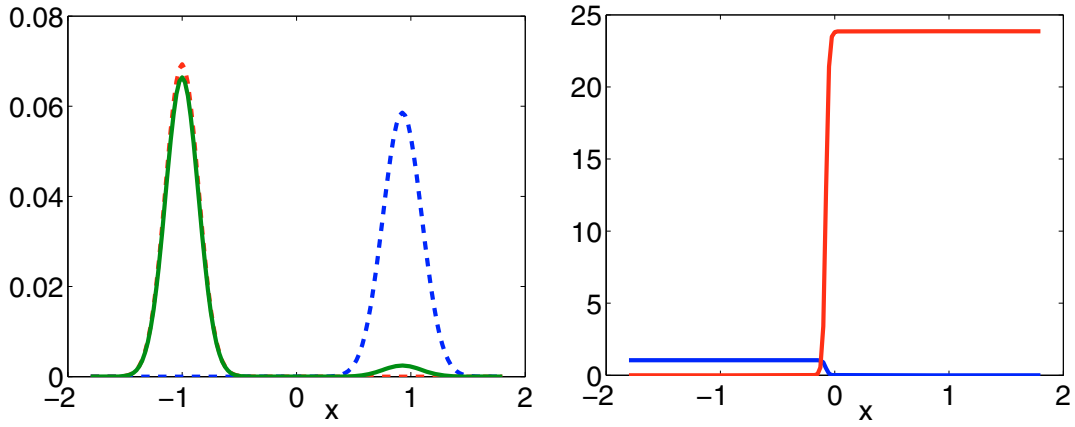

FIGURE 4. Left: invariant measures μ_1 (dashed), μ_2 (dashed) and the full measure μ (solid) as computed from the DNS sampling with $\tau = 0.5$. Right: corresponding approximate invariant densities Φ_1 and Φ_2.

We now aim at understanding the full kernel $K_t(x, x_0)$ as an superposition kernel $k_t(x, x_0)$ in the sense of Section 3.1:

$$\mu(x)K_t(x, x_0)\mu(x_0) \approx \mu(x)k_t(x, x_0)\mu(x_0) = \sum_{j=1}^{2} \alpha_j \mu_j(x)k_j(x, x_0)\mu_j(x_0),$$

where the k_i are the two Gaussian kernels that have been determined above, and

$$\alpha_1 = 0.948, \qquad \alpha_2 = 0.052,$$

as a further result of the Gaussian mixture model. But this means that we determined everything that is required to construct the superposition kernel $k_t(x, x_0)$ that approximates $K_t(x, x_0)$. In terms of the weighted 1-norm $\|k_t - K\|_{1,\mu}$ the agreement is very good (as far as this can be checked based on histograms of k_t).

Based on these results we are able to inspect the approximate invariant densities Φ_1 and Φ_2 (see Fig. 4) and compute the corresponding overlap $O_{12} = \langle \Phi_1, \Phi_2 \rangle_\mu$. We get

$$O_{12} = 2.8 \times 10^{-9},$$

such that we can enter the above metastability results with $\epsilon = O_{12} = 2.8 \times 10^{-9}$. This shows that the kernel is ϵ-metastable, but it also shows that the current approximation contains "too much metastability" in the following sense: the metastable sets exhibit stronger metastability with respect to the approximate superposition kernel than with respect to the original process. This is a consequence of the fact that the approximation of the sampling kernel was based on the 1-norm instead on the 2-norm (see Rem. 3.8); in the 1-norm the weights of the transition regions are small since the sampling kernels are small there.

5.2. **Assignment to metastable and transition states**

Let us consider the above example again. Let us denote the available time series generated by DNS by $x_{0:m} = \{x_0, \ldots, x_m\}$. From this we get the time series underlying the sampling kernel; this will be denoted $z_{0:m-1} = \{z_0, \ldots, z_{m-1}\}$, where $z_n = (x_n, x_{n+1})$.

According to the previous analysis we have two main metastable states. Points in the shortened time series $x_{0:m-1}$ can be *assigned* to these states *via* the almost invariant densities Φ_i, $i = 1, 2$, in the sense of constructing the two sets

$$\mathcal{M}_i = \{x_n : 0 \leq n \leq m-1,\ \Phi_i(x_n) > \theta \|\Phi_i\|_\infty\}, \quad i = 1, 2,$$

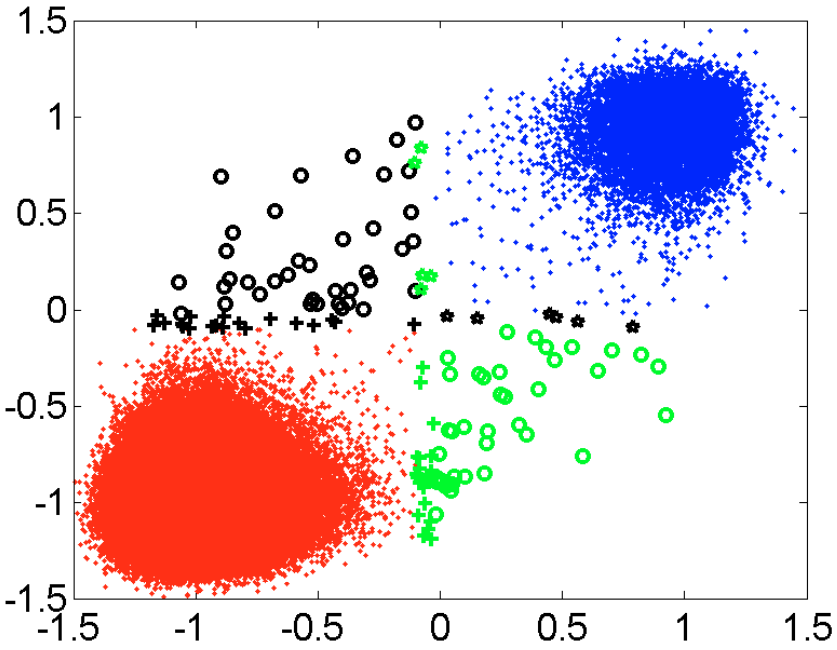

FIGURE 5. Sampling of the sampling kernel with coloring of points according to assignment as described in the text ($\theta = 0.95$). The two clouds of grey and black dots represent $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The crosses indicate M_{10} (black) and M_{01} (green), the circles M_{12} (black) and M_{21} (light grey), and the stars M_{20} (black) and M_{02} (light grey).

where $\theta > 0.5$ is some appropriate user-selected threshold (*e.g.*, 0.95). The properties of the Φ_i guarantee that $\mathcal{M}_1 \cap \mathcal{M}_2 = \emptyset$. All other x_n will be collected in the *transition set*

$$\mathcal{M}_0 = \{x_n : 0 \le n \le m-1,\ \Phi_i(x_n) \le \theta \|\Phi_i\|_\infty,\ i = 1, 2\}.$$

Transitions are events n where $x_n \in \mathcal{M}_j$ for $j = 0, 1, 2$ but $x_{n+1} \notin \mathcal{M}_j$. We can classify these *via* the timeseries z as follows: First introduce

$$M_{ij} = \{z_n = (x_n, x_{n+1}) :\ 0 \le n \le m-1,\ x_n \in \mathcal{M}_i \text{ and } x_{n+1} \in \mathcal{M}_j\}.$$

See Figures 5 and 6 for illustrations of these sets for different values of θ.

Let now $\#A$ denote the number of elements in the set A. Then, we observe that for all $i = 0, 1, 2$

$$\#\mathcal{M}_i = \sum_j \#M_{ij},$$

and the optimal Markov transition matrix (in a MLE sense, *i.e.*, under the condition of the observation $\hat{X}$ made) between the sets $\mathcal{M}_i$, $i = 0, 1, 2$, has transition probabilities

$$p(i, j) = \frac{\#M_{ij}}{\#\mathcal{M}_i}.$$

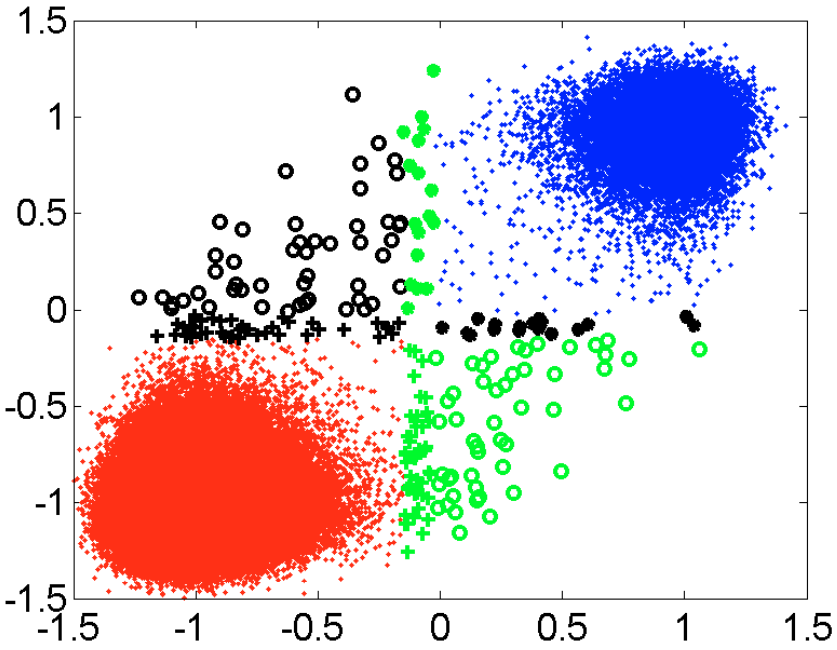

FIGURE 6. Sampling of the sampling kernel with coloring of points according to assignment as described in the text ($\theta = 0.99$). The two clouds of grey and black dots represent $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The crosses indicate M_{10} (black) and M_{01} (light grey), the circles M_{12} (black) and M_{21} (light grey), and the stars M_{20} (black) and M_{02} (light grey).

In our case we get ($\theta = 0.95$)

$$\top = \begin{pmatrix} 0 & 0.8112 & 0.1888 \\ 0.0001 & 0.9997 & 0.0002 \\ 0.0003 & 0.0030 & 0.9967 \end{pmatrix}.$$

The eigenvalues of this matrix are $1.0000, 0.9965, -0.0001$ which illustrates that, as long as we are interested in metastability, the process can be further aggregated by means of the PCCA algorithm [8,9] into the 2×2 process

$$\top = \begin{pmatrix} 0.9998 & 0.0002 \\ 0.0031 & 0.9969 \end{pmatrix},$$

which describes the jumps between $\mathcal{M}_1$ and $\mathcal{M}_2$.

Let us conclude our considerations with the following sketch of the algorithmic approach we are advocating herein:

- Use Algorithm 1 to find an approximate superposition kernel $\sum_{i=1}^{M} \alpha_i \kappa_i$ based on a sampling (time series with lag time τ) of the sampling kernel of the system under investigation.
- Construct the almost invariant functions Φ_i based on the invariant measures of the approximate superposition kernel and compute the transition matrix $\top$ as outlined in this section.
- Take the Markov chain associated with $\top$ as description of the effective dynamics of the system on timescale τ and the dynamics associated to the sampling kernels κ_i, $i = 1, \ldots, M$, as local dynamics within each of the M metastable sets.

In Section 2.3 of [32] this approach has been applied to the small peptide trialanine; it has been demonstrated that its results coincide with results of other algorithmic approaches to metastable dynamical behavior in molecular systems.

6. Relation to other approaches

6.1. **Extended state space**

Let us consider again Markov kernels k_i, $i \in I = \{1, \ldots, N\}$, with invariant probability measures μ_i, and associated flat space transition densities $\rho_i(x, \cdot)$. Assume that k_i and ρ_i are associated with a lag time τ. Let us assume that the measures μ_i are absolutely continuous with respect to the Lebesgue measure and that the ρ_i are positive (almost) everywhere. Let the underlying state space be Ω. Now we consider the extension of the process to the extended state space $\Omega \times I$, *i.e.*, the number of the respective component k_i of the process now is part of the state information. Now, consider the following Markov transition density on this extended state space:

$$\rho_{\text{ext}}(x, i, y, j) = \rho_i(x, y) T_{ij}(y), \tag{6.1}$$

where $i, j \in I$, and T_{ij} denotes the i, j-entry of the transition matrix $T = \exp(\tau R)$ of the Markov jump process with rate matrix R on I. R and T are supposed to depend on y such that $\pi = (\pi_j(y))_{j \in I}$, the invariant measure of $T(y)$, satisfies

$$\pi_j(y) = \alpha_j \mu_j(y) / \mu(y),$$

with some fixed positive numbers α_i that do not depend on y and satisfy $\sum_{j \in I} \alpha_j = 1$, and $\mu(y) = \sum_j \alpha_j \mu_j(y)$. Under these conditions we easily verify that the invariant measure of the extended transition function with density ρ_{ext} is

$$\mu_{\text{ext}}(x, i) = \alpha_i \mu_i(x).$$

We can now lump the extended transition function together again, *i.e.*, we consider its *marginal transition density*

$$\rho_{\text{mar}}(x, y) = \sum_{i,j \in I} \pi_i(x) \rho_{\text{ext}}(x, i, y, j). \tag{6.2}$$

Interestingly we then get back to the transition density of the superposition process on Ω:

$$\rho_{\text{mar}}(x, y) = \sum_{i \in I} \alpha_i \frac{\mu_i(x)}{\mu(x)} \rho_i(x, y).$$

6.2. **Towards HMMSDE**

Let ρ_i, $i = 1, \ldots, N$, be Markov transition densities and $R \in \mathbb{R}^{N \times N}$ a rate matrix. The transition matrix associated with R and step $t = 1$ is $\top = \exp(R)$. Let π be the invariant measure of $\top$, *i.e.*, $\pi^T \top = \pi^T$. Consider the extended state space $\hat{\Omega} = \Omega \times \{1, \ldots, N\}$. Then introduce the transition density

$$\rho(x, i, y, j) = \rho_i(x, y) \top_{ij}$$

which defines a "1-step" Markov kernel on $\hat{\Omega}$. In contrast to what has been considered above we do no longer assume that $\top$ depends on the target state $y \in \Omega$.

In case that the $\rho_i(x, y)$ are transition functions of Ornstein-Uhlenbeck processes

$$\dot{x} = -DV^{(i)}(x) + \Sigma^{(i)} \dot{W},$$

the such defined process is governed by the HMMSDE model [16]

$$\begin{aligned} \dot{x}(t) &= -DV^{(i(t))}(x(t)) + \Sigma^{(i(t))} \dot{W}, \\ i(t) &= \text{Markov jump process with rate matrix } R. \end{aligned}$$

Concerning parameterization of this process by the time series at hand, we get the ρ_i by approximation of the sampling kernel and the transition matrix from the above EM algorithm and counting scheme. In [16–18], another approach to parameter estimation for the HMMSDE model has been presented. Further investigations will have to work out whether these algorithms can also be used for the advanced kernel approximation scheme.

Appendix A

As an example in a *deterministic* setting, we derive an explicit formula for the transition kernel in the case of a partially observed Hamiltonian system. Consider the ordinary differential equation $\dot{z} = F(z)$ with flow map Φ^t such that $z(t) = \Phi^t z_0$ if $z(0) = z_0$. Now assume that the trajectory is only partially observed, *i.e.*, instead of the full state $z = (x, \xi)$ we do only consider the part $x = Qz$ where Q denotes the projection from the state space onto the subspace corresponding to x. Then, the observed process has the form $x(t) = Q\Phi^t(x_0, \xi_0)$. Furthermore assume that the flow map Φ^t leaves the measure $\pi = \mu \otimes \nu$ invariant, of which we assume that it is absolutely continuous with respect to the Lebesgue measure and decomposes according to $\pi(x, \xi) = \mu(x)\nu(\xi)$. Under these conditions the transfer operator of the observed process $x(t)$ on time scale τ has the following form [29,30] in the function space L^2_μ:

$$P^\tau f(x) = \int f(Q\Phi^{-\tau}(x, \xi))\, \nu(\xi)\, \mathrm{d}\xi.$$

Rewriting it in the above notation under the assumption that the μ is almost everywhere positive exhibits that the associated kernel in L^2_μ has the form

$$k_\tau(x, y) = \frac{1}{\mu(y)} \int \delta(y - Q\Phi^{-\tau}(x, \xi))\, \nu(\xi)\, \mathrm{d}\xi,$$

where δ denotes the usual delta distribution which is used here for the sake of simplicity.

In order to understand what kind of function k_τ may be, consider the following scenario which originates from molecular dynamics applications: there $\dot{z} = F(z)$ should be thought of as a Hamiltonian system with position x, momentum ξ, and Hamiltonian $H(z) = T(\xi) + V(x)$. Hence, $F(z) = -JDH(z)$ where DH denotes the derivative of H with respect to $z = (x, \xi)$, and J is the typical skew-symmetric block matrix $J = [0, I;\ -I, 0]$. The associated flow then leaves the measure

$$\pi(x, \xi) = \mu(x)\nu(\xi) = \frac{1}{Z_x} \exp(-\beta V(x)) \cdot \frac{1}{Z_\xi} \exp(-\beta T(\xi))$$

invariant, where Z_x and Z_ξ are appropriate normalization constants and β is some arbitrary positive number. In this context, all above assumptions are satisfied for arbitrary potential energies V and kinetic energies T that grow strong enough. In order to allow a glimpse on the structure of k_τ let us specifically choose the case of a one-dimensional position coordinate x and $V(x) = x^2/2$ and $T(\xi) = \xi^2/2$, and τ so that $s = \sin(\tau) \neq 0$. Then $Q\Phi^{-\tau}(x_0, \xi_0) = \cos(\tau)x_0 - \sin(\tau)\xi_0$, and we find (set $c = \cos(\tau)$)

$$k_\tau(x, y) = \frac{1}{\mu(y)} \frac{1}{s} \nu\Big(\frac{cx - y}{s}\Big).$$

This then results in a sampling kernel of Gaussian form

$$\kappa_\tau(x, y) = \frac{1}{sZ_xZ_\xi} \exp\left(-\frac{\beta}{2}\left[x^2 + \left(\frac{cx - y}{s}\right)^2\right]\right) = \frac{1}{sZ_xZ_\xi} \exp\left(-\frac{\beta}{2}\left[\frac{x^2 - 2cxy + y^2}{s^2}\right]\right).$$

Next, consider the Hamiltonian $H(x, \xi) = T(\xi) + V(x)$ where $V(x)$ now denotes the double well potential shown in Figure 2 (*cf.* Sect. 5.1). Let $\beta = 5$ and $\tau = 0.5$. After performing a DNS of the projected Hamiltonian

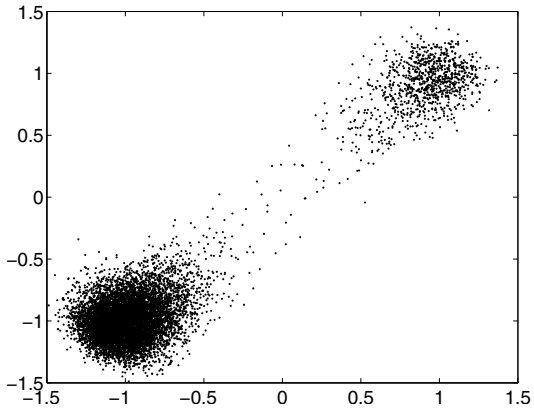

FIGURE 7. Sampling of the sampling kernel for projected Hamiltonian system as described in the text.

system (with Verlet discretization in time with timestep 0.005), we *directly* get a sampling of the associated sampling kernel $\kappa_\tau(x, x_0)$ with respect to the invariant measure μ of the process. This sampling is shown in Figure 7. We observe that the sampling kernel can well be approximated by a superposition of two Gaussian.

REFERENCES

[1] M. Belkin and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in *Advances in Neural Information Processing Systems* **14**, T.G. Diettrich, S. Becker and Z. Ghahramani Eds., MIT Press (2002) 585–591.
[2] J. Bilmes, *A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* ICSI-TR-97-021 (1997).
[3] A. Bovier, M. Eckhoff, V. Gayrard and M. Klein, Metastability in stochastic dynamics of disordered mean-field models. *Probab. Theor. Rel. Fields* **119** (2001) 99–161.
[4] J. Chodera, N. Singhal, V. Pande, K. Dill and W. Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Comp. Chem.* **126** (2007) 155101.
[5] E.B. Davies, Metastable states of symmetric Markov semigroups I. *Proc. London Math. Soc.* **45** (1982) 133–150.
[6] M. Dellnitz and O. Junge, On the approximation of complicated dynamical behavior. *SIAM J. Numer. Anal.* **36** (1999) 491–515.
[7] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B* **39** (1977) 1–38.
[8] P. Deuflhard and M. Weber, Robust Perron cluster analysis in conformation dynamics. *Lin. Alg. App.* **398** (2005) 161–184.
[9] P. Deuflhard, W. Huisinga, A. Fischer and Ch. Schütte, Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.* **315** (2000) 39–59.
[10] R. Duda, P. Hart and D. Stork, *Pattern Classification.* Wiley (2001).
[11] L. Elsner and S. Friedland, Variation of the Discrete Eigenvalues of Normal Operators. *P. Am. Math. Soc.* **123** (1995) 2511–2517.
[12] S. Fischer, B. Windshügel, D. Horak, K.C. Holmes and J.C. Smith, Structural mechanism of the recovery stroke in the myosin molecular motor. *Proc. Natl. Acad. Sci. USA* **102** (2005) 6873–6878.
[13] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach and Ch. Schütte, Identification of biomolecular conformations from incomplete torsion angle observations by Hidden Markov Models. *J. Comp. Chem.* **28** (2007) 1384–1399.
[14] H. Frauenfelder, S.G. Sligar and P.G. Wolynes, The energy landscapes and motions of proteins. *Science* **254** (1991) 1598–1603.
[15] H.O. Hartley, Maximum likelihood estimation from incomplete data. *Biometrics* **14** (1958) 174–194.
[16] I. Horenko and Ch. Schütte, Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics. *Multiscale Model. Simul.* **7** (2008) 731–773.
[17] I. Horenko, E. Dittmer, A. Fischer and Ch. Schütte, Automated model reduction for complex systems exhibiting metastability. *Mult. Mod. Sim.* **5** (2006) 802–827.
[18] I. Horenko, C. Hartmann, Ch. Schuette and F. Noé, Data-based parameter estimation of generalized multidimensional Langevin processes. *Phys. Rev. E* **76** (2007) 016706.

[19] W. Huisinga and B. Schmidt, Metastability and dominant eigenvalues of transfer operators, in *Advances in Algorithms for Macromolecular Simulation*, C. Chipot, R. Elber, A. Laaksonen, B. Leimkuhler, A. Mark, T. Schlick, C. Schütte and R. Skeel Eds., *Lect. Notes Comput. Sci. Eng.* **49**, Springer (2005) 167–182.
[20] W. Huisinga, S. Meyn and Ch. Schütte, Phase transitions and metastability in Markovian and molecular systems. *Ann. Appl. Probab.* **14** (2004) 419–458.
[21] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M.E. Bowman, J. Noel, M. Gruebele and J. Kelly, Structure-function-folding relationship in a ww domain. *Proc. Natl. Acad. Sci. USA* **103** (2006) 10648–10653.
[22] S. Lafon and A.B. Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning and data set parameterization. *IEEE Trans. Pattern Anal. Mach. Intell.* **28** (2006) 1393–1403.
[23] B.B. Laird and B.J. Leimkuhler, Generalized dynamical thermostating technique. *Phys. Rev. E* **68** (2003) 016704.
[24] B. Nadler, S. Lafon, R.R. Coifman and I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.* **21** (2006) 113–127.
[25] F. Noé, D. Krachtus, J.C. Smith and S. Fischer, Transition networks for the comprehensive characterization of complex conformational change in proteins. *J. Chem. Theory Comput.* **2** (2006) 840–857.
[26] A. Ostermann, R. Waschipky, F.G. Parak and G.U. Nienhaus, Ligand binding and conformational motions in myoglobin. *Nature* **404** (2000) 205–208.
[27] L.R. Rabiner, A tutorial on HMMs and selected applications in speech recognition. *Proc. IEEE* **77** (1989).
[28] Ch. Schütte and W. Huisinga, On conformational dynamics induced by Langevin processes, in *EQUADIFF 99 – International Conference on Differential Equations* **2**, B. Fiedler, K. Gröger and J. Sprekels Eds., World Scientific (2000) 1247–1262.
[29] Ch. Schütte and W. Huisinga, Biomolecular conformations can be identified as metastable sets of molecular dynamics, in *Handbook of Numerical Analysis* **X**, P.G. Ciarlet and C. Le Bris Eds., Elsevier (2003) 699–744.
[30] Ch. Schütte, A. Fischer, W. Huisinga and P. Deuflhard, A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys., Special Issue on Computational Biophysics* **151** (1999) 146–168.
[31] Ch. Schütte, W. Huisinga and P. Deuflhard, Transfer operator approach to conformational dynamics in biomolecular systems, in *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, B. Fielder Ed., Springer (2001) 191–223.
[32] C. Schütte, F. Noe, E. Meerbach, P. Metzner and C. Hartmann, Conformations dynamics, in *Proceedings of ICIAM 2007*, Section on Public Talks (to appear).
[33] G. Singleton, Asymptotically exact estimates for metastable Markov semigroups. *Quart. J. Math. Oxford* **35** (1984) 321–329.
[34] D. Wales, *Energy Landscapes*. Cambridge University Press, Cambridge (2003).