Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

MARCO SARICH[1], RALF BANISCH[1], CARSTEN HARTMANN[1] AND
CHRISTOF SCHÜTTE[1,2]

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

[2]*Zuse Institute Berlin, Germany*

Markov State Models for Rare Events in Molecular Dynamics

ZIB-Report 13-52 (August 2013)

Herausgegeben vom
Konrad-Zuse-Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Markov State Models for Rare Events in Molecular Dynamics

Marco Sarich[1], Ralf Banisch[1], Carsten Hartmann[1] and Christof Schütte[1,2]

[1] *Department of Mathematics and Computer Science, Freie Universität Berlin, Germany*
[2] *Zuse Institute Berlin, Germany*

Abstract

Rare but important transition events between long lived states are a key feature of many molecular systems. In many cases the computation of rare event statistics by direct molecular dynamics (MD) simulations is infeasible even on the most powerful computers because of the immensely long simulation timescales needed. Recently a technique for spatial discretization of the molecular state space designed to help overcome such problems, so-called Markov State Models (MSMs), has attracted a lot of attention. We review the theoretical background and algorithmic realization of MSMs and illustrate their use by some numerical examples. Furthermore we introduce a novel approach to using MSMs for the efficient solution of optimal control problems that appear in applications where one desires to optimize molecular properties by means of external controls.

Keywords. rare events; Markov state models; long timescales; optimal control

MSC 2010 subject classifications. 49M29, 60J28, 60J35, 93E20

1 Introduction

Stochastic processes are widely used to model physical, chemical or biological systems. The goal is to approximately compute interesting properties of the system by analyzing the stochastic model. As soon as randomness is involved there are mainly two options for performing this analysis: (1) Direct sampling and (2) the construction of a discrete coarse grained model of the system. In a direct sampling approach, one tries to generate a statistically significant amount of events that characterize the property of the system one is interested in. For this purpose, computer simulations of the model are a powerful tool. For example, an event could refer to the transition between two well-defined macroscopic states of the system. In chemical applications such transitions can often be interpreted as reactions, or in the context of a molecular system as conformational changes. Interesting properties are e.g. average waiting times for such reactions or conformational changes and along which pathways the transitions typically occur. The problem with a direct sampling approach is that many interesting events are so called rare events. Therefore the computational effort for generating sufficient statistics for reliable estimates is very high, and, particularly if the state space is continuous and high dimensional, estimation by direct numerical simulation is infeasible.

Available techniques for rare event simulations in continuous state space are discussed in [1]. In this article, we will discuss approach (2) to the estimation of rare event statistics via *discretization* of the state space of the system under consideration. That is, instead of dealing with the computation of rare events for the original, continuous process, we will approximate them by a so-called Markov State Model (MSM) with *discrete* finite state space. The reason is that for such a discrete model one can numerically compute many interesting properties without simulation, mostly by solving linear systems of equations as in discrete transition path theory (TPT) [2]. We will see that this approach, called Markov State Modelling, *avoids*

the combinatorial explosion of the number of discretization elements with increasing size of the molecular system in contrast to other methods for spatial discretization.

The actual construction of an MSM requires to sample certain transition probabilities of the underlying dynamics between sets. The idea is (1) to choose the sets such that the samling effort is much lower than the direct estimation of the rare events under consideration, and (2) to compute all interesting quantities for the MSM from its transition matrix, cf. [2,3]. There are many examples for the successful application of this strategy. In [4], for example, it was used to compute dominant folding pathways for the PinWW domain in explicit solvent. However, we have to make sure that the Markov State Model approximates the orginial dynamics well enough. For example, the MSM should correctly reproduce the timescales of the processes of interest. These approximation issues have been discussed since more than a decade now [5, 6]; in this article we will review the present state of research on this topic. In the algorithmic realization of Markov State Modelling for realistic molecular systems the transition probabilities and the respective statistical uncertainties are estimated from short MD trajectories only, cf. [7]. This makes Markov State Modelling applicable to many different molecular systems and processes, cf. [8–13].

In the first part of this article we will discuss the approximation quality of two different types of Markov State Models that are defined with respect to a full partition of state space or with respect to so-called core sets. We will also discuss the algorithmic realization of MSMs and provide references to the manifold of realistic applications to molecular systems in equilibrium that are available in the literature today.

The second part will show how to use MSMs for optimizing particular molecular properties. In this type of application one wants to steer the molecular system at hand by external controls in a way such that a pre-selected molecular property is optimized (minimized or maximized). That is, one wants to compute a specific external control from a family of admissible controls that optimizes the property of interest under certain side conditions. The property to be optimized can be quite diverse: For example, it can be (1) the population of a certain conformation that one wants to maximize under a side condition that limits the total work done by the external control or (2) the mean first passage time to a certain conformation that one wants to minimize (in order to speed up a rare event) but under the condition that one can still safely estimate the mean first passage time of the uncontrolled system. The theoretical background of case (1) has been considered in [14], for example, and of case (2) in [1, 15]. There one finds the mathematical problem that has to be solved in order to compute the optimal control. Here we will demostrate that one can use MSMs for the efficient solution of such a mathematical problem (for both cases). We will see that the spatial discretization underlying an MSM turns the high-dimensional continuous optimal control problem into a rather low-dimensional discrete optimal control problem of the same form that can be solved efficiently. Based on these insights, MSM discretization yields an efficient algorithm for solving the optimal control problem whose performance we will outline in some numerical examples including an application to Alanine dipeptide.

2 MSM Construction

Let $(X_t)_{t\geq 0}$ be a time-continuous Markov process on a continuous state space E, e.g. $E \subset R^d$. That is, X_t is the state of the molecular system at time t as resulting from any usually used from of molecular dynamics simulation, be it based on Newtonian dynamics with thermostats or resulting from Langevin dynamics or other diffusion molecular dynamics models. The idea of Markov State Modelling is to derive a Markov chain $(\hat{X}_k)_{k\in\mathbb{N}}$ on a finite and preferably small state space $\hat{E} = \{1, ..., n\}$ that models characteristic dynamics of the continuous process

(X_t). For example, in molecular dynamics applications such characteristic dynamics could refer to protein folding processes [16, 17], conformational rearrangements between native protein substates [18, 19], or ligand binding processes [20]. Since the approximating Markov chain $(X_k)_{k\in\mathbb{N}}$ lives on a finite state space, the construction of an MSM boils down to the computation of its transition matrix P

$$P_{ij} = \mathbb{P}[\hat{X}_{k+1} = j | \hat{X}_k = i]. \tag{1}$$

The main benefit is that for a finite Markov chain one can compute many interesting dynamical properties directly from its transition matrix, e.g. timescales and metastability in the system [5, 21, 22], a hierarchy of important transition pathways [2], or mean first passage times between selected states. With respect to na MSM, these computations should be used afterwards to answer related questions for the original continuous process. To do this we must be able to link the states of the Markov chain back to spatial information of the original process and the approximation of the process (X_t) by the MSM must be valid in some sense.

Having this in mind the first natural idea is to let the states of an MSM correspond to sets $A_1, ..., A_n \subset E$ in continuous state space that form a full partition, i.e.

$$A_i \cap A_j = \emptyset \text{ for } i \neq j, \qquad \bigcup_{i=1}^{n} A_i = E. \tag{2}$$

Typical choices for such sets are box discretizations or voronoi tesselations [23]. For such a full partition it is trivial to also define a corresponding discretized process by the original switching dynamics between the sets. For a given lag time $\tau > 0$, we can define the index process

$$\tilde{X}_k = i \Leftrightarrow X_{k\tau} \in A_i. \tag{3}$$

It is well known that this process is not Markovian, mainly due to the so called recrossing problem. It refers to the fact that the original process typically crosses the boundary between two sets A_i and A_j several times when transitions take place, as illustrated in Figure 1. This results in cumulative transitions between indices i and j for the index process, that is, a not memoryless transition behavior.

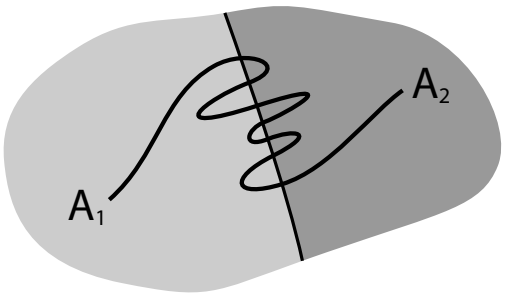


Figure 1: Cumulative transitions between two sets along boundaries are typical.

The non-Markovianity of the index process is often seen as a problem in Markov State Modeling because many arguments assume that $\tilde{X}_k$ is a Markov process. In this article, we

will *not* make this assumption. We interpret the process $(\tilde{X}_k)$ as a tool to construct the following transition matrix P^τ

$$P^\tau_{ij} = \mathbb{P}[\tilde{X}_{k+1} = j | \tilde{X}_k = i] = \mathbb{P}[X_{(k+1)\tau} \in A_j | X_{k\tau} \in A_i] \tag{4}$$

and hence the MSM as the Markov chain $(\hat{X}_k)_{k\in\mathbb{N}}$ associated with this transition matrix. From above it is clear that in general we have $\hat{X}_k \neq \tilde{X}_k$ and in [24] it was analyzed how these two processes relate in terms of density propagation. In the following, we will show under which assumptions and in which sense the MSM $(\hat{X}_k)$ will be a good approximation of the original dynamics given by (X_t). For convenience we will usually write $P^\tau \equiv P$ and leave the τ-dependence implicit.

3 Analytical Results

In order to compare the MSM to the continuous process we introduce one of the key objects for our analysis, the *transfer operator* of a Markov process. We assume that the Markov process (X_t) has a unique, positive invariant probability measure μ and that it is time-reversible. Then, for any time-step $t \geq 0$ we define the transfer operator T_t via the property

$$\int_A T_t v(y)\mu(dy) = \int_E v(x)p(t,x,A)\mu(dx) \qquad \text{for all measurable } A \tag{5}$$

as an operator $T_t : L^2(\mu) \to L^2(\mu)$. Here, $p(t,x,A) = \mathbb{P}[X_t \in A | X_0 = x]$ defines the transition probability measure and $L^2(\mu)$ denotes the Hilbert space of functions v with

$$\int_E v(y)^2 \mu(dy) \leq \infty \tag{6}$$

and the scalar product

$$\langle v, w\rangle = \int_E v(y)w(x)\mu(dy). \tag{7}$$

Note that T_t is nothing else than the propagator of densities under the dynamics, but the densities are understood as densities with respect to the measure μ. That is, if the Markov process is initially distributed according to

$$\mathbb{P}[X_0 \in A] = \int_A v_0(x)\mu(dx), \tag{8}$$

its probability distribution at time t is given by

$$\mathbb{P}[X_t \in B] = \int_B v_t(x)\mu(dx), \qquad v_t = T_t v_0. \tag{9}$$

The benefit of working with μ-weighted densities is that the transfer operator T_t becomes essentially self-adjoint on $L^2(\mu)$ for all cases of molecular dynamics satisfying some form of detailed balance condition. Hence, it has real eigenvalues and orthogonal eigenvectors with respect to (7) (or at least the dominant spectral elements are real-valued). Moreover, the construction of an MSM can be seen as a projection of the transfer operator [25] . Assume Q is an orthogonal projection in $L^2(\mu)$ onto an n-dimensional subspace $D \subset L^2(\mu)$ with $\mathbb{1} \in D$, and $\chi_1, ..., \chi_n$ is a basis of D. Then, the so called projected transfer operator $QT_\tau Q : D \to D$ has the matrix representation

$$P_Q = PM^{-1}, \tag{10}$$

with the non-negative, invertible mass matrix $M \in \mathbb{R}^{n,n}$ with entries

$$M_{ij} = \frac{\langle \chi_i, \chi_j \rangle}{\langle \chi_i, \mathbb{1} \rangle}. \tag{11}$$

The matrix $P \in \mathbb{R}^{n,n}$ is also non-negative and has entries

$$P_{ij} = \frac{\langle \chi_i, T_\tau \chi_j \rangle}{\langle \chi_i, \mathbb{1} \rangle}. \tag{12}$$

Full Partition MSM. If we choose $\chi_i = \mathbb{1}_{A_i}$ to be the characteristic function of set A_i for $i = 1, ..., n$, one can easily check that we get $M = I$ to be the identity matrix and

$$P_{ij} = \mathbb{P}_\mu[X_\tau \in A_j | X_0 \in A_i] \tag{13}$$

as in (4). The subscript μ shall indicate that $X_0 \sim \mu$. So the transition probabilities are evaluated along equilibrium paths.

The previously constructed transition matrix of the MSM based on a full partition can be interpreted as a projection onto a space of densities which are constant on the partitioning sets. This interpretation of an MSM is useful since it allows to analyze its approximation quality. For example, in [25, 26] it is proven that we can reproduce an eigenvalue λ of a self-adjoint transfer operator T_t by the MSM by choosing the subspace appropriately. That is, if u is a corresponding normalized eigenvector, Q the orthogonal projection to a subspace D with $\mathbb{1} \in D$, then there exists an eigenvalue $\hat{\lambda}$ of the projected transfer operator QT_tQ with

$$|\lambda - \hat{\lambda}| \leq \lambda_1 \delta (1 - \delta^2)^{-\frac{1}{2}},$$

where $\lambda_1 < 1$ is the largest non-trivial eigenvalue of T_t and $\delta = \|u - Qu\|$.

In particular, for $\delta \leq \frac{3}{4}$ one can simplify the equation to

$$|\lambda - \hat{\lambda}| \leq 2\lambda_1 \delta. \tag{14}$$

An eigenvalue λ_i of the transfer operator directly relates to an implied timescales $\mathcal{T}_i$ of the system via

$$\mathcal{T}_i = -\frac{\tau}{\log(\lambda_i)}. \tag{15}$$

So the transition matrix (4) that we construct from transitions between the sets $A_1, ..., A_n$ will generate a Markov chain that will reproduce the original timescales well if the partitioning sets are chosen such that the corresponding eigenvectors are almost constant on these sets. In this case $\delta = \|u - Qu\|$, that is the approximation error of the eigenvector by a piecewise constant function on the sets will be small.

The projection error δ depends on our choice of the discretizing sets. As an example let us consider a diffusion in the potential that is illustrated in Figure 2, that is, the reversible Markov process given by the stochastic differential equation

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\varepsilon} dB_t, \tag{16}$$

where V is the potential, B_t denotes a Brownian motion and $\varepsilon > 0$.

The figure also shows a choice of three sets that form a full partition of state space. The computation of the transition matrix (4) for $\sigma = 0.7$ and a lag time $\tau = 1$ yields

$$P_Q = P = \begin{pmatrix} 0.9877 & 0.0123 & 0.0000 \\ 0.0420 & 0.9160 & 0.0419 \\ 0.0000 & 0.0123 & 0.9877 \end{pmatrix}$$

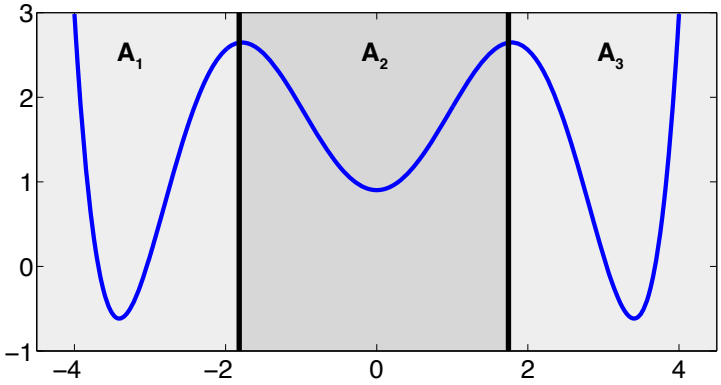


Figure 2: A potential with three wells and a choice of 3 sets A_1, A_2, A_3.

that has three eigenvalues $\lambda_0 = 1, \lambda_1 = 0.9877, \lambda_2 = 0.9037$. The following table shows the two resulting implied timescales (15) in comparison to the timescales of the original system.

	T_1	T_2
original	103.7608	11.9566
full partition 3 sets	80.6548	9.8784

As one can see, the timescales are strongly underestimated. This is a typical phenomenon. From a statistical point of view, the recrossing problem will lead to cumulatively appearing transition counts when one computes the transition probabilities $\mathbb{P}_\mu[X_\tau \in A_j | X_0 \in A_i]$ from a trajectory (X_t), as discussed above. Therefore on average transitions between sets seem to become too likely and hence the processes in the coarse grained system get accelerated. We have seen in (14) that this cannot happen if the associated eigenvectors can be approximated well by the subspace that corresponds to the MSM. Figure 3 shows the first non-trivial eigenvector u_1 belonging to the timescale $T_1 = 103.7608$ and its best-approximation by a step function.

The eigenvector is indeed almost constant in the vicinity of the wells, but within the transition region between the wells the eigenvector is varying and the approximation by a step function is not accurate. So we have two explanations why the main error is introduced in the region close to shared boundaries of neighboring sets: (1) because of recrossing issues and (2) because of the main projection error of the associated eigenvector. Of course, one solution would be an adaptive refinement of the discretization, that is, one could choose a larger number of smaller sets such that the eigenvector is better approximated by a step function on these sets. In the following section, we will present an alternative solution for overcoming the recrossing problem and reducing the projection error without refining the discretization.

4 The Core Set Approach

From (10) we know how to compute a matrix representation for a projected transfer operator for an arbitrary subspace $D \subset L^2(\mu)$. For a given basis $\chi_1, ..., \chi_n$ we have to compute (11)

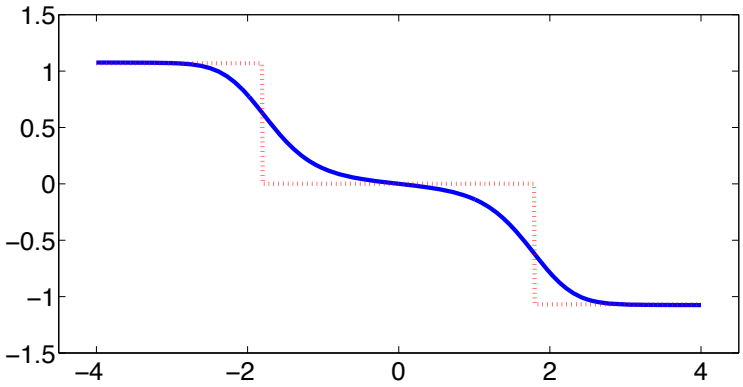


Figure 3: The first non-trivial eigenvector u_1 (solid blue) and its projection Qu_1 (dashed red) onto step functions that are constant on A_1, A_2, A_3.

and (12), so

$$M_{ij} = \frac{\langle \chi_i, \chi_j \rangle}{\langle \chi_i, \mathbb{1} \rangle}, \qquad P_{ij} = \frac{\langle \chi_i, T_\tau \chi_j \rangle}{\langle \chi_i, \mathbb{1} \rangle}. \tag{17}$$

In general, the evaluation of these scalar products for arbitrary basis functions is a non-trivial task. On the other hand, we have seen that for characteristic functions $\chi_i = \mathbb{1}_{A_i}$ on a full partition we do not have to compute the scalar products numerically since the matrix entries have a stochastic interpretation in terms of transition probabilities between sets (13). This means they can be directly estimated from a trajectory of the process which is a strong computational advantage, particularly in high dimensional state spaces.

Now, the question is if there is another basis than characteristic functions that a) is more adapted to the eigenvectors of the transfer operator, and b) still leads to a probabilistic interpretation of the matrix entries (17) such that scalar products never have to be computed. The basic idea is to stick to a set-oriented definition of the basis, but to relax the full partition constraint. We will define our basis with respect to so called *core sets* $C_1, ..., C_n \subset E$ that are still disjoint, so $C_i \cap C_j = \emptyset$, but they do not have to form a full partition. Figure 4 suggests that this could lead to a reduction of the recrossing phenomenon since the sets do not share boundaries anymore.

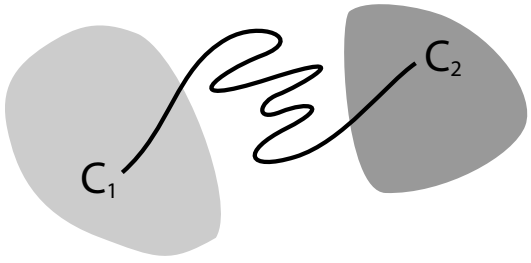


Figure 4: Core sets do not have to share boundaries anymore. This can reduce the recrossing effect.

Now, we use the core sets to define our basis functions $\chi_1, ..., \chi_n$. Assume T_τ is again

a self-adjoint transfer operator and consider n core sets $C_1, ..., C_n$. For every i, take the committor function χ_i of the process with respect to core set C_i, that is, $\chi_i(x)$ denotes the probability to hit the core set C_i next rather than the other core sets when starting the process in x. If we now study the the projection Q onto the space spanned by these committor functions, the two following properties hold [25, 27].

(P1) The matrices M and P in (10) can be written as

$$M_{ij} = \mathbb{P}_\mu[\tilde{X}_k^+ = j | \tilde{X}_k^- = i], \qquad P_{ij} = \mathbb{P}_\mu[\tilde{X}_{k+1}^+ = j | \tilde{X}_k^- = i], \tag{18}$$

where $(\tilde{X}_k^+)$ and $(\tilde{X}_k^-)$ are forward and backward milestoning processes [25, 28], that is, $\tilde{X}_k^- = i$ if the process came at time $t = k\tau$ last from core set C_i and $\tilde{X}_k^+ = j$ if the process went next to core set C_j after time $t = k\tau$.

(P2) Let u_i be an eigenvector of T_τ that is almost constant on the core sets. Let the region $C = E \backslash \bigcup_i C_i$ that is not assigned to a core set be left quickly enough, so $\mathbb{E}_x[\tau(C^c)] \ll \mathcal{T}_i$ for all $x \in C$, where $\mathcal{T}_i$ is the timescale associated with u_i and $\mathbb{E}_x[\tau(C^c)]$ is the expected hitting time of $C^c = \bigcup_i C_i$ when starting in $x \in C$. Then, $\|u_i - Qu_i\|$ is small, so the committor approximation to the eigenvector is accurate.

The message behind (P1) is that it is possible to relax the full partition constraint and use a core set discretization that does not cover the whole state space. We can still define a basis for a projection of the transfer operator that leads to a matrix representation that can be interpreted in terms of transition probabilities.

Important remark: The construction of the projection onto the committors is only necessary for theoretical purposes. In practice, neither the committor functions, nor scalar products between the committors have to be computed numerically, since the matrix entries of M and P can be estimated from trajectories again.

Property (P2) yields that the relaxation of the full partition constraint should also lead to an improvement of the MSM if the region C between the core sets is typically left on a faster timescale than the processes of interest take place. Let us get back to the example from above. We will see that we can achieve a strong improvement of the approximation by simply excluding a small part of state space from our discretiazion. In Figure 5 we have turned our initial full partition into a core set discretization by removing parts of the transition region between the wells.

The matrix $P_Q = PM^{-1}$ that represents the projection $QT_\tau Q$ of the transfer operator onto the committor space associated with the core sets is given by

$$P_Q = \begin{pmatrix} 0.9897 & 0.0103 & 0.0000 \\ 0.0352 & 0.9298 & 0.0351 \\ 0.0000 & 0.0103 & 0.9897 \end{pmatrix}$$

Comparing to the MSM for the full partition one can see that transitions between indices i and j, $i \neq j$ are less likely. As the following table shows this leads to a far more accurate reproduction of the timescales in the system.

	T_1	T_2
original	103.7608	11.9566
3 core sets	100.8066	11.9145
full partition 3 sets	80.6548	9.8784

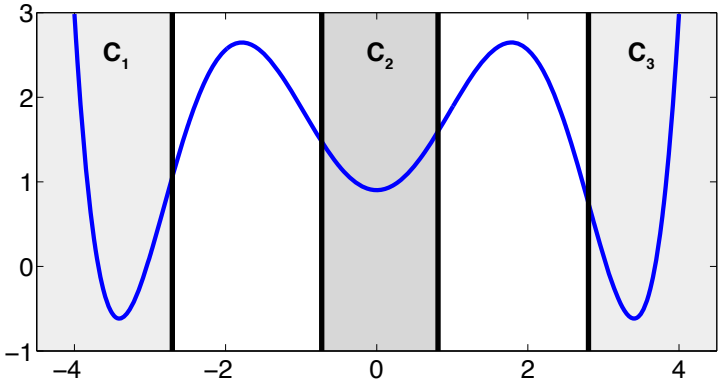


Figure 5: Excluding a small region of state space from the sets $A_1, A_2, A3$ as in Figure 2 to form core sets C_1, C_2, C_3 that do not share boundaries anymore.

From the discussion above this has to be expected because the eigenvectors are almost constant in the vicinity of the wells and we removed a part of state space from the discretization that is typically left quickly compared to the timescales T_1 and T_2. So, the committor functions should deliver a good approximation of the first two eigenvectors. Figure 6 underlines this theoretical result.

5 Practical Considerations and MD Applications

In the previous sections we have interpreted the construction of an MSM as projection of the dynamics onto some finite dimensional ansatz space. We have discussed two types of spaces that both have been defined on the basis of a set discretization. First, we chose a full partition of state space and the associated space of step functions, and second we analyzed a discretization by core sets and the associated space spanned by committor functions. These two methods have the advantage that the resulting projections lead to transition matrices for the MSM with entries that are given in terms of transition probabilities between the sets. That is, one can compute estimates for the transition matrices from simulation data. This is an important property for practical applications because it means that we never need to compute committor functions, or scalar products between committors or step functions. We rather generate trajectories $x_0, x_1, ...x_N$ of the process (X_t), let us say for a time step $h > 0$, so $x_i = X_{hi}$. For example, we can then define for a full partition $A_1, ..., A_m$ and a lag time $\tau = nh$ the discrete trajectory $s_k = i \Leftrightarrow x_k \in A_i$ and compute the matrix $\hat{P}$

$$\hat{P}_{ij} = \frac{C_{ij}}{\sum_j C_{ij}}, \qquad C_{ij} = \sum_{k=0}^{N-n} \mathbb{1}_{\{s_k=i\}} \mathbb{1}_{\{s_{k+n}=j\}}. \tag{19}$$

It is well-known [29] that $\hat{P}$ is a maximum likelihood estimator for the full partition MSM transition matrix (4). Similarly one can also compute estimates for a core set MSM by using the definition of milestoning processes [27, 28]. That is, if we have core sets $C_1, ..., C_m$, a lag time $\tau = nh$ as before, and we define discrete milestoning trajectories by

$$s_k^- = i \Leftrightarrow x_k \in A_i \text{ or came last from } A_i \text{ before time } k$$
$$s_k^+ = i \Leftrightarrow x_k \in A_i \text{ or went next to } A_i \text{ after time } k,$$

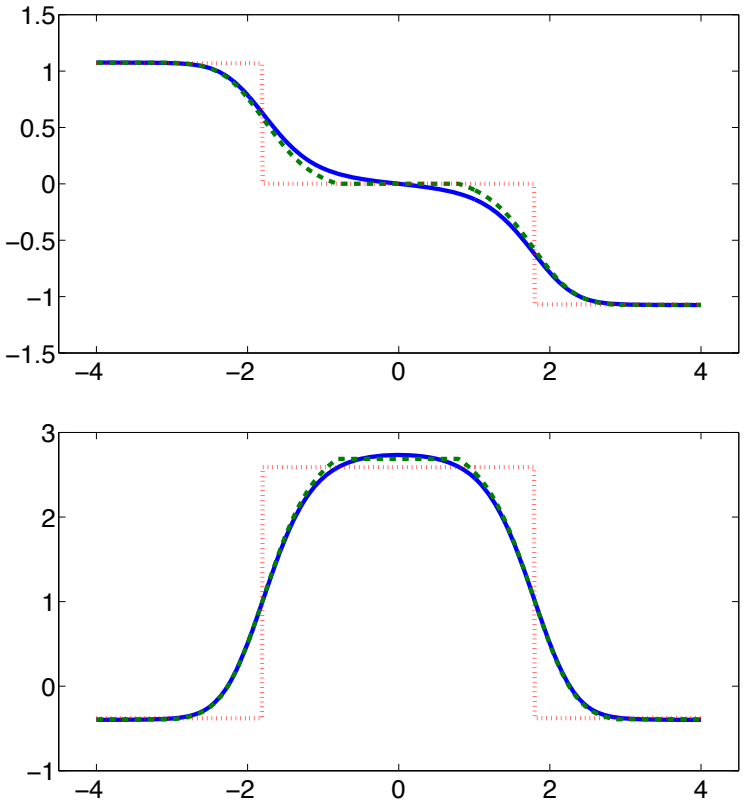


Figure 6: Upper panel: The first non-trivial eigenvector u_1 (solid blue) and its projection $Q_f u_1$ (finely dashed red) onto stepfunctions (full partition) and its projection $Q_c u_1$ (dashed green) onto committors (core sets). Lower panel: The same plot for the second non-trivial eigenvector u_2.

we can compute an estimator $\hat{P}_Q = \hat{P}\hat{M}^{-1}$ of the core set MSM matrix (10) by counting transitions:

$$\hat{P}_{ij} = \frac{C_{ij}}{\sum_j C_{ij}}, \qquad C_{ij} = \sum_{k=0}^{N-n} \mathbb{1}_{\{s_k^- = i\}} \mathbb{1}_{\{s_{k+n}^+ = j\}}, \tag{20}$$

$$\hat{M}_{ij} = \frac{N_{ij}}{\sum_j N_{ij}}, \qquad N_{ij} = \sum_{k=0}^{N} \mathbb{1}_{\{s_k^- = i\}} \mathbb{1}_{\{s_k^+ = j\}}. \tag{21}$$

Since in practice we will only have a finite amount of data available, we will have statistical errors when constructing an MSM. This is an additional error to the projection error related to the discretization that we have discussed above. On the other hand, one should note that these errors are *not independent* of each other. For example, it is clear that if we take a full partition of state space and we let the partition become arbitrarily fine by letting the number of sets go to infinity, the discretization error will vanish. At the same time, for a fixed amount of statistics, the statistical error will become arbitrarily large because we will need to compute more and more estimators for transition events between the increasing number of sets. For more information on statistical errors we refer to the literature [29, 30].

Besides the choice of discretization and the available statistics, the estimates above also depend on a lag time τ. This dependence can be used to validate an MSM by a Chapman Kolmogorov test [29]. This is based on the fact that the MSM matrices approximately form a semi-group for all large enough lag times $\tau > \tau^*$, although for small lag times this is typically not true due to memory effects. These facts also motivate to look at something like an infinitesimal generator that approximately generates these MSM transition matrices for large enough lag times. In [27], two types of generator constructions have been compared for a core set setting. The first generator K is simply constructed from the transition rates between the core sets in the milestoning sense, that is

$$K_{ij} = \lim_{T\to\infty} \frac{N_{ij}^T}{R_i^T}, i \neq j \qquad K_{ii} = -\sum_{j\neq i} K_{ij}, \tag{22}$$

where N_{ij}^T is the amount of time in $[0,T]$ the process has spent on its way from core set C_i to C_j, and R_i^T is the total time in $[0,T]$ the process came last from C_i. On the other hand, one can see [27, 31] that $K^* = KM^{-1}$ with the mass matrix M from above (18) can be interpreted as a projection of the original generator of the process, and also as derivative of the core set MSM from above, i.e.

$$K^* = \lim_{\tau\to 0} \frac{PM^{-1} - I}{\tau}, \tag{23}$$

where P depends on τ (17).

Let us now analyze how the choice of core sets, particularly the size of the core sets, influences the resulting approximation. Therefore, we consider an MD example that was discussed in [27], namely one molecule of alanine dipeptide monitored via its ϕ and ψ backbone dihedral angles. Two core sets are defined as balls with radius r around the two points with angular coordinates $x_\alpha = (-80, -60)$ and $x_\beta = (-80, 170)$. The stationary distribution of the process and the two centers of the core sets x_α, x_β in the angular space are shown in Figure 7.

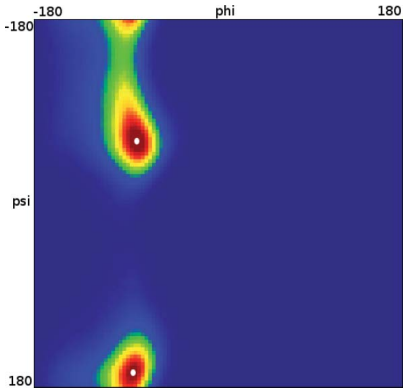


Figure 7: The stationary distribution of alanine dipeptide and the two centers of the core sets x_α, x_β in the angular space as white dots.

For computing a reference timescale several MSMs based on full partitions using 10,15, and 250 sets have been constructed for increasing lag times. In [27], it is shown that in each setting the estimate for the longest implied timescale of the process converged to ≈ 19 ps

for large enough τ. Now the implied timescales for the two different generators K (22) and K^* (23) are computed. In Figure 8, the resulting timescales are plotted against the reference timescale ≈ 19 ps for varying size of the core sets.

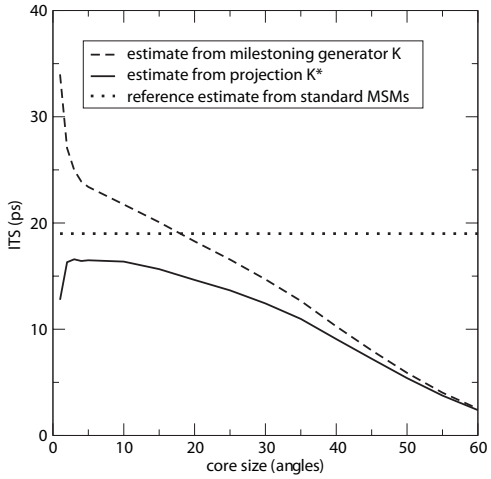


Figure 8: Estimate of the implied timescales from K (22), the projected generator K^* (23) and the reference computed from several full partition MSMs.

One can see that the estimate by the milestoning generator K is rather sensitive to the size of core sets. It overestimates the timescales for small core sizes and underestimates it for larger core sizes. On the other hand, the projected genetator K^* can never overestimate the timescale due to its interpretation as projection. It is also rather robust against the choice of size of the core sets until the core sets become too large, e.g. $r > 15$. Then, the discretization becomes close to a full partition discretization using only two sets. In this case the timescales have to be underestimated heavily because of recrossing phenomena. On the other hand, the underestimation for very small core sets has to be explained by a lack of statistics. When the core sets are chosen arbitrarily small, it is clearly more difficult for the process to hit the sets and therefore transition events become rare. Note that for the straightforward milestoning generator K the processes seem to become very slow, but for the projected generator $K^* = KM^{-1}$ this effect is theoretically corrected by the mass matrix M. Nevertheless, in both cases the generation of enough statistics will be problematic for too small core sets.

Further Applications in MD. Markov State Modelling has been show to apply successfully to many different molecular systems like peptides including time-resolved spectroscopic experiments [10–12], proteins and protein folding [4, 9, 13], or DNA [32]. In most of the respective publications full partition MSMs are used and the underlying dicretization is based on cluster finding methods, see [29] for a review. Core set based approaches have been used just recently [10, 27].

Very rare transitions between discretization sets. When constructing a full partition or a core set MSM we have to estimate transition probabilities between sets in state space, and it can happen that we cannot avoid that some of these transitions are *very* rare. That is,

the transition probabilities for a lag time τ between some sets can be non-zero, but small even if compared to the remaining transition probabilities that are small already. This is why it is important to note that neglecting these very rare transitions during the construction of an MSM does *not* harm its approximation quality. For example, we can define for a transition matrix P another transition matrix $\tilde{P}$ by

$$\tilde{P}_{ij} = \begin{cases} P_{ij}, & i \neq j, (i,j) \notin R \\ 0, & i \neq j, (i,j) \in R \end{cases} \tag{24}$$

where R denotes the set of pairs of indices for which the transition are very rare and for which we set the transition probability to zero. If the Markov chain is reversible and $(i,j) \in R \Leftrightarrow (j,i) \in R$, one can show that for all ordered eigenvalues $\lambda_k(P)$ and $\lambda_k(\tilde{P})$ it holds that

$$|\lambda_k(P) - \lambda_k(\tilde{P})| \leq \max_i \sum_{j \neq i, (i,j) \in R} P_{ij}. \tag{25}$$

That is, if we cannot estimate a very small transition probability P_{ij} for a very rare transition event between two sets A_i and A_j, and even totally neglect this probability by setting it to zero, the timescales of the MSM remain almost unaffected. Thus, if we compute the set of "first order" transition probability of a system correctly enough and ignore all "higher order" ones, then our accuracy will not be spoiled.

6 MSM for Optimal Control Problems

In this section we will borrow ideas from the previous section and explain how MSMs can be used to discretize optimal control problems that are linear-quadratic in the control variables and which appear in e.g. sampling of rare events. Specifically, we consider the case that $(X_t)_{t \geq 0}$ is the solution of

$$dX_t = (\sqrt{2}u_t - \nabla V(X_t))dt + \sqrt{2\varepsilon}dB_t\,, \tag{26}$$

with potential V, Brownian motion B_t and temperature $\varepsilon > 0$ as in (16) and an unknown control variable $u\colon [0,\infty) \to \mathbb{R}^d$ that is chosen so as to minimize the cost function

$$J(u;x) = \mathbb{E}\left[\int_0^\tau \left(f(X_s) + \frac{1}{2}|u_t|^2\right) ds \middle| X_0 = x\right]. \tag{27}$$

(The factors of $1/2$ and $\sqrt{2}$ in front of the control terms are for notational convenience.) Here $f \geq 0$ is a bounded continuous function called *running cost* and $\tau < \infty$ (a.s.) is a random stopping time that is determined by X_t hitting a given target set $A \subset E$, i.e. $\tau = \inf\{t > 0\colon X_t \in A\}$, in other words, we are interested in controlling $X_t = X_t^u$ until it reaches A. As an example, consider the case $f = 1$ and $A = C_1$ with the potential considered in Figure 5, which amounts to the situation that one seeks to minimize the time to reach the core set C_1 by tilting the potential towards the target set C_1; tilting the potential too much is prevented by the quadratic penalization term in the cost functional that grows when too much force is applied.

Other choices of f in (26) result in alternative applications. One obvious application would be to set $\tau = T$ to a fixed time and f to the characteristic function of the complement of a conformation set C, $f = \mathbb{1}_{E \setminus C}$. In this case, minimization of J wrt. the control u_t would mean maximization of the probability to find the system in the conformation C until time T

under a penalty on the external work done to the system. See [14] for more details on such applications.

There are other types of cost functions J one might consider, e.g. control until a deterministic finite time $\tau = T$ is reached, or even $\tau \to \infty$, and the construction would follow analogously. For compactness we consider here only cost functions as in (27).

Optimal control and equilibrium expectation values. It turns out that when minimizing J it is sufficient to consider control strategies that are Markovian and depend only on X_t, i.e. we consider feedback laws of the form $u_t = \alpha(X_t)$ for some smooth function $\alpha\colon E \to \mathbb{R}^d$. Moreover only controls with finite energy are considered, for otherwise $J(u;x) = \infty$. For control problems of the form (26)–(27) the optimal feedback function can be shown to be $\alpha^*(x) = -\sqrt{2}\nabla W$ where W is the value function or optimal-cost-to-go [1,15]

$$W(x) = \min_u J(u;x) \tag{28}$$

with the minimum running over all admissible Markovian feedback strategies. It can be shown that W satisfies the following dynamic programming equation of Hamilton-Jacobi-Bellman type (see [33]):

$$\begin{aligned} LW(x) - |\nabla W(x)|^2 + f &= 0 \\ W|_A &= 0\,, \end{aligned} \tag{29}$$

with the second-order differential operator

$$L = \varepsilon\Delta - \nabla V \cdot \nabla$$

that is the infinitesimal generator of the process X_t for $u = 0$. If the value function W is known, it can be plugged into the equation of motion which then turns out to be of the form

$$dX_t^* = -\nabla U(X_t^*)dt + \sqrt{2\varepsilon}dB_t\,, \tag{30}$$

with the new potential

$$U(x) = V(x) + 2W(x)\,.$$

The difficulty is that equation (29) is a nonlinear partial differential equation and for realistic high-dimensional systems it is not at all obvious how to discretize it, employing any kind of state space partitioning. It has been demonstrated in [14, 15] that (29) can be transformed into a linear equation by a logarithmic transformation. Setting $W(x) = -\varepsilon \log \phi(x)$ it readily follows, using chain rule and equation (29), that ϕ solves the linear equation

$$\begin{aligned} (L - \varepsilon^{-1} f)\phi &= 0 \\ \phi|_A &= 1\,. \end{aligned} \tag{31}$$

The last equation is linear and can be solved by using MSMs as we will show below. Moreover, by the Feynman-Kac theorem [34], the solution to (31) can be expressed as

$$\phi(x) = \mathbb{E}\left[\exp\left(-\frac{1}{\varepsilon}\int_0^\tau f(X_t)dt\right)\middle| X_0 = x\right], \tag{32}$$

where X_t solves the control-free equation

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\varepsilon}dB_t\,.$$

That is, the optimal control for (26) can be computed by solving (31) which can be done *in principle* via Monte-Carlo approximation of the expected value in (32) if critical slowing down by rare events can be avoided.

Remark. The optimization problem (28) admits an interpretation in terms of entropy minimization: Let $Q = Q_x^u$ and $P = Q_x^0$ denote the path probability measures of controlled and uncontrolled trajectories starting at x at time $t = 0$, and set

$$Z = \int_0^\tau f(X_s)\, ds,$$

then it follows that we can write

$$W(x) = \min_{Q \ll P} J(u; x), \qquad J(u; x) = \int \left\{ Z + \varepsilon \log\left(\frac{dQ}{dP}\right) \right\} dQ, \tag{33}$$

where the notation "$Q \ll P$" means that Q has a density[1] with respect to P. It turns out that for every such Q there is exactly one control strategy u such that $Q = Q_x^u$ is generated by (26), in this sense the notation in (33) is meaningful. The second term

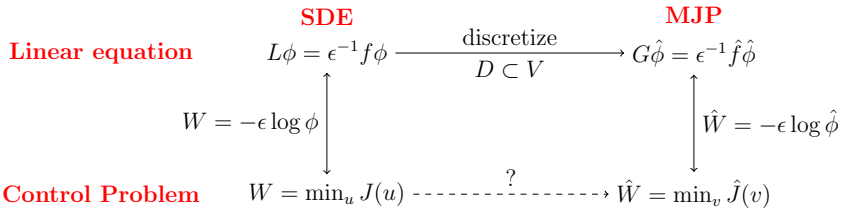
$$H(Q\|P) = \varepsilon \int \log\left(\frac{dQ}{dP}\right) dQ$$

is the relative entropy or Kullback-Leibler divergence between Q and P. For details on this matter that are based on Girsanov transformations for stochastic differential equations we refer to [35] or the article [1] in this special issue.

7 MSM Discretization of Optimal Control Problems

The basic idea is now to choose a subspace $D \subset L^2(\mu)$ with basis $\chi_1, \ldots, \chi_n$ as in Markov state modelling and then discretize the dynamic programming equation (29) of our optimal control problem by projecting the equivalent log transformed equation (31) onto that subspace. As we will see the resulting discrete matrix equation can be transformed back into an optimal control problem for a discrete Markov jump process (MJP).

We will do this construction for the full partition case $\chi_i = \mathbb{1}_{A_i}$ and the core set case $\chi_i = q_i$ discussed earlier. We will see that in both cases, we arrive at a structure-preserving discretization of the original optimal control problem where the states of the corresponding MJP will be related to the partition subsets A_i. The first case will give us back a well-known lattice discretization for continuous control problems, the Markov chain approximation [36]. This is illustrated in the following diagram:



Subspace projection. The key step for the discretization is that we pick a suitable subspace $D \subset L^2(\mu)$ that is adapted to the boundary value problem (31). Specifically, we require that the subspace contains the constant function $\mathbb{1} \in D$ and that it gives a good representation of the most dominant metastable sets. To this end we choose basis functions $\chi_1, \ldots, \chi_{n+1}$ with the following properties:

[1] That is, the density function dQ/dP exists, is almost everywhere positive and normalized.

(S1) The χ_i form a partition of unity, that is $\sum_{i=1}^{n+1} \chi_i = \mathbb{1}$.

(S2) The χ_i are adapted to the boundary conditions in (31), that is $\chi_{n+1}|_A = 1$ and $\chi_i|_A = 0$ for $i \in \{1, \ldots, n\}$.

Now let Q be the orthogonal projection onto D, and define the matrices

$$F_{ij} = \frac{\langle \chi_i, f\chi_j \rangle}{\langle \chi_i, \mathbb{1} \rangle}, \qquad K_{ij} = \frac{\langle \chi_i, L\chi_j \rangle}{\langle \chi_i, \mathbb{1} \rangle}.$$

Now, if ϕ solves the linear boundary value problem (31), then the coefficients $\hat{\phi}_1, \ldots, \hat{\phi}_{n+1}$ of its finite-dimensional representation $Q\phi = \sum_j \hat{\phi}_j \chi_j$ on the subspace D satisfy the constrained linear system

$$\begin{aligned} \sum_{j=1}^{n+1} \left(K_{ij} - \varepsilon^{-1} F_{ij} \right) \hat{\phi}_j = 0\,, \quad i \in \{1, \ldots, n\} \\ \hat{\phi}_{n+1} = 1\,, \end{aligned} \tag{34}$$

that is the discrete analogue of (31). The discrete solution $\hat{\phi} = Q\phi$ is optimal in the sense of being the best approximation of ϕ in the energy norm, i.e.,

$$\|\phi - \hat{\phi}\|_A = \inf_{\psi \in D} \|\phi - \psi\|_A\,, \tag{35}$$

where

$$\|\phi\|_A^2 = \langle \phi, (\varepsilon^{-1} f - L)\phi \rangle$$

is the energy norm on $L^2(\mu)$, and the infimum runs over all functions $\psi \in L^2(\mu)$ that are of the form $\psi(x) = \sum_j \psi_j \chi_j(x)$ with coefficients $\psi_j \in \mathbb{R}$. This is a standard result about projections of PDEs, see [37] for details.[2] In analogy with equation (14) we can use the above result to get the error estimate

$$\|\phi - \hat{\phi}\|_\mu^2 \leq \left(1 + \frac{1}{\delta^2} \|QAQ^\perp\|^2 \right) \inf_{\psi \in D} \|\phi - \psi\|_\mu^2 \tag{36}$$

where $A = \varepsilon^{-1} f - L$ is a shorthand for the operator appearing in (31) and the constant $\delta > 0$ is defined such that $\|v\|_A^2 \geq \delta \|v\|_\mu^2$ holds for all $v \in L^2(\mu)$; see [38]. The bottom line of (35) is that discretizing (31) via (34) minimizes the projection error measured in the energy norm. Since all functions are μ-weighted, the approximation will be good in regions visited with high probability and less good in regions with lower probability. The error estimate (36) is along the lines of the MSM approximation result: If we switch to the norm on $L^2(\mu)$, the function $\hat{\phi} = Q\phi$ is still almost the best approximation of ϕ, provided that A leaves the subspace D almost invariant. As was pointed out earlier this is exactly the case when the χ_i are close to the eigenfunctions of A (e.g., when the system is metastable).

The best approximation error $\|Q^\perp \phi\|_\mu = \inf_{\psi \in D} \|\phi - \psi\|_\mu$ which appears in (36) will vanish if the χ_i form an arbitrarily fine full partition of E. If we follow the core set idea from section 4 and choose the χ_i to be committor functions, we have good control over $\|Q^\perp \phi\|_\mu$. Due to [38],

$$\|Q^\perp \phi\|_\mu \leq \|P^\perp \phi\|_\mu + \mu(C)^{1/2} \left[\kappa \|f\|_\infty + 2\|P^\perp \phi\|_\infty \right], \tag{37}$$

[2] By the same argument as in the previous sections $A = \varepsilon^{-1} f - L$ is symmetric and positive definite as an operator on the weighted Hilbert space $L^2(\mu)$. Moreover $\|\phi\|_A^2 = \varepsilon^{-1} \langle \phi, f\phi \rangle + \varepsilon \langle \nabla\phi, \nabla\phi \rangle$.

where $C = E \backslash \cup_i C_i$ is the transition region, $\kappa = \sup_{x \in C} \mathbb{E}_x \tau_{E \backslash C}$ is the maximum expected time of hitting the metastable set from outside (which is short) and P is the orthogonal projection onto the subspace $V = \{v \in L^2(\mu), v = const \text{ on every } C_i\} \subset L^2(\mu)$. Note that $P^\perp \phi = 0$ on C. The errors $\|P^\perp \phi\|_\mu$ and $\|P^\perp \phi\|_\infty$ measure how constant the solution ϕ is on the core sets. Hence (37) together with (36) gives us complete control over the approximation error of our projection method even if we consider just a few core sets. In section 8, we will investigate the full and core set partition cases further.

Properties of the projected problem. We introduce now the diagonal matrix Λ with entries $\Lambda_{ii} = \sum_j F_{ij}$ (zero otherwise) and the full matrix $G = K - \varepsilon^{-1}(F - \Lambda)$, and rearrange (34) as follows:

$$\begin{aligned} \sum_{j=1}^{n+1} \left(G_{ij} - \varepsilon^{-1} \Lambda_{ij} \right) \hat{\phi}_j = 0 \,, \quad i \in \{1, \ldots, n\} \\ \hat{\phi}_{n+1} = 1 \,, \end{aligned} \tag{38}$$

This equation can be given a stochastic interpretation. To this end let us introduce the vector $\pi \in \mathbb{R}^{n+1}$ with nonnegative entries $\pi_i = \langle \chi_i, \mathbb{1} \rangle$ and notice that $\sum_i \pi_i = 1$ follows immediately from the fact that the basis functions χ_i form a partition of unity, i.e. $\sum_i \chi_i = \mathbb{1}$. This implies that π is a probability distribution on the discrete state space $\hat{E} = \{1, \ldots, n+1\}$. We summarise properties of the matrices K, F and G, see also [38]:

(M1) K is a generator matrix of a MJP $(\hat{X}_t)_{t \geq 0}$ (i.e., K is a real-valued square matrix with row sum zero and positive off-diagonal entries) with stationary distribution π that satisfies detailed balance

$$\pi_i K_{ij} = \pi_j K_{ji} \,, \quad i, j \in \hat{E}$$

(M2) $F \geq 0$ (entry-wise) with $\pi_i F_{ij} = \pi_j F_{ji}$ for all $i, j \in \hat{E}$.

(M3) G has row sum zero and satisfies $\pi^T G = 0$ and $\pi_i G_{ij} = \pi_j G_{ji}$ for all $i, j \in \hat{E}$; furthermore there exists a constant $0 < C < \infty$ such that $G_{ij} \geq 0$ for all $i \neq j$ if $\|f\|_\infty \leq C$. In this case equation (38) admits a unique and strictly positive solution $\hat{\phi} > 0$.

It follows that if the running costs f are such that (M3) holds, then G is a generator matrix of a MJP that we shall denote by $(\hat{X}_t)_{t \geq 0}$, and (38) has a unique and positive solution. In this case the logarithmic transformation $\hat{W} = -\varepsilon \log \hat{\phi}$ is well-defined. It was shown in [39] that $\hat{W}$ can be interpreted as the value function of a Markov decision problem with cost functional (cf. also [33])

$$\hat{J}(v; i) = \mathbb{E} \left[\int_0^\tau \left(\hat{f}(\hat{X}_s) + k(\hat{X}_s, v_s) \right) ds \middle| \hat{X}_0 = i \right] \tag{39}$$

that is minimized over the set of Markovian control strategies $v \colon \hat{E} \to (0, \infty)$ subject to the constraint that the controlled process $\hat{X}_t = \hat{X}_t^v$ is generated by G^v where

$$G_{ij}^v = \begin{cases} v(i)^{-1} G_{ij} v(j) \,, & i \neq j \\ -\sum_{j \neq i} G_{ij}^v \,, & i = j \end{cases} \tag{40}$$

with stopping time $\tau = \inf\{t > 0 \colon \hat{X}_t = n+1\}$ and running costs

$$\hat{f}(i) = \Lambda_{ii} \,, \quad k(i, v) = \varepsilon \sum_{j \neq i} G_{ij} \left\{ \frac{v(j)}{v(i)} \left[\log \frac{v(j)}{v(i)} - 1 \right] + 1 \right\}. \tag{41}$$

Properties of the projected problem, cont'd. From [39] we know that the optimal cost

$$\hat{W}(i) = \min_v \hat{J}(v; i)$$

is given by $\hat{W} = -\epsilon \log \hat{\phi}$ where $\hat{\phi}$ solves (38), with the optimal feedback strategy given by $v^*(i) = \hat{\phi}_i$ (see [33]). We list additional properties:

(i) The v-controlled system has the unique invariant distribution

$$\pi^v = (\pi_1^v, \dots, \pi_{n+1}^v), \quad \pi_i^v = \frac{v(i)^2 \pi_i}{Z_v}$$

with Z_v an appropriate normalization constant; in terms of the value function $\pi^* = \pi^{v^*}$ reads

$$\pi^* = (\pi_1^*, \dots, \pi_{n+1}^*), \quad \pi_i^* = \frac{1}{Z_*} e^{-2\varepsilon^{-1}\hat{W}(i)} \pi_i.$$

(ii) G^v is reversible and stationary with respect to π^v, i.e., $\pi_i^v G_{ij}^v = \pi_j^v G_{ji}^v$ for all $i, j \in \hat{E}$.

(iii) $\hat{J}$ admits the same interpretation as (33) in terms of the relative entropy:

$$\hat{W}(i) = \min_{Q \ll P} \hat{J}(v; i), \qquad \hat{J}(v; i) = \int \left\{ \hat{Z} + \varepsilon \log \left(\frac{dQ}{dP} \right) \right\} dQ$$

where P denotes expectation with respect to the uncontrolled MJP $\hat{X}_t$ starting at $\hat{X}_0 = i$, Q denotes the path measure of the corresponding controlled process with generator G^v and

$$\hat{Z} = \int_0^\tau \hat{f}(\hat{X}_s)\, ds\,.$$

A few remarks seem in order: Item (i) of the above list is in accordance with the continuous setting, in which the optimally controlled dynamics is governed by the new potential $U = V + 2W$ and has the stationary distribution $\mu^* \propto \exp(-2\epsilon^{-1}W)\mu$ with μ being the stationary distribution of the uncontrolled process. Hence the effect of the control on the invariant distribution is the same in both cases. Further note that optimal strategies change the jump rates according to

$$G_{ij}^{v^*} = G_{ij} e^{-\varepsilon^{-1}\left(\hat{W}(j) - \hat{W}(i)\right)}, \tag{42}$$

that is $\hat{W}$ acts as an effective potential as in the continuous case, and the change in the jump rates can be interpreted in terms of Kramer's law for this effective potential.

This completes our derivation of the discretized optimal control problem, and we now compare it with the continuous problem we started with for the case of a full partition of E and a core set partition of E.

8 Markov Chain Approximations and Beyond

Full partitions. Let E be fully partitioned into disjoint sets $A_1, \dots, A_{n+1}$ with centers $x_1, \dots, x_{n+1}$ and such that $A_{n+1} := A$, and define $\chi_i := \chi_{A_i}$. These χ_i satisfy the assumptions (S1) and (S2) discussed in section 7. Since they are not overlapping, F is diagonal, and

$$\hat{f}(i) = \frac{1}{\pi_i} \int_{A_i} f(x)\mu(x)dx = \mathbb{E}_\mu[f(X_t)|X_t \in A_i] \tag{43}$$

is just obtained by averaging $f(x)$ over the cell A_i. (43) is also a sampling formula for $\hat{f}(i)$. It follows directly that $G = K$, and in particular (M3) holds for any f. One can show that K has components

$$K_{ij} \approx \frac{1}{\Delta_{ij}} e^{-\beta(V(\bar{x}_{ij})-V(x_i))}, \quad \Delta_{ij}^{-1} = \beta^{-1} \frac{m(S_{ij})}{m(h_{ij})m(A_i)} \tag{44}$$

if i and j are neighbours ($K_{ij} = 0$ otherwise). Here m is the Lebesgue measure, and h_{ij}, S_{ij} and $\bar{x}_{ij}$ are defined as in Figure 9. K is the generator of a MJP on the cells A_i and coincides with the so-called *finite volume approximation* of L discussed in [40]. It is reversible with stationary distribution

$$\pi_i = \int_{A_i} d\mu \approx m(A_i)e^{-\beta V(x_i)}.$$

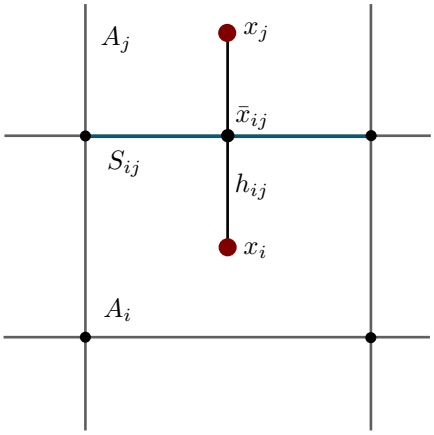


Figure 9: The mesh for the full partition.

One can show that the approximation error vanishes for $n \to \infty$. K and π can be computed from the potential V and the geometry of the mesh. By inspecting (12) and (13), we see that K is connected to the transition matrix P^τ of a full partition MSM with lagtime τ by

$$\lim_{\tau\to 0} \frac{1}{\tau}\left(P_{ij}^\tau - M_{ij}\right) = \lim_{\tau\to 0} \frac{1}{\pi_i}\langle \chi_i, \frac{1}{\tau}(T_\tau - \mathbb{1})\chi_j\rangle = \frac{1}{\pi_i}\langle \chi_i, L\chi_j\rangle = K_{ij},$$

thus K is the generator of the semigroup of transition matrices P^τ. Therefore we could obtain K by sampling in the same way we obtained P^τ through equation (19) in section 5. This is difficult however due to recrossing problems for small τ, see e.g. [41]. Finally, let us note in passing that we can drastically simplify k^v if the cells A_i are boxes of length h. Denote the elementary lattice vectors by e_n. Then

$$k^v(i) = \frac{1}{2}|u^v(i)|^2 + \mathcal{O}(h), \quad u_n^v(i) := \frac{1}{\sqrt{2}} \frac{\epsilon}{2h} \left(\log v(i+e_n) - \log v(i-e_n)\right)$$

which establishes the connection to the continuous case. But more is true: The whole discrete control problem reduces to first order in h to the well-known Markov chain approximation (MCA) [36], which allows us to use convergence theory for MCAs to conclude that for $n \to \infty$, optimal control and value function of the discrete control problem converge to their continuous counterparts. More details can be found in [38].

Core set partition. Now we choose core sets $C_1, \ldots, C_{n+1}$ with $C_{n+1} = A$ and we let $\chi_i = q_i$ to be the commitor function of the process with respect to C_i as in section 4. These χ_i satisfy the assumptions (S1) and (S2) discussed in section 7. The projection onto the committor basis also allows for a stochastic interpretation. Recall the definition of the forward and backward milestoning process $\tilde{X}_t^{\pm}$ from (18). The discrete costs can be written as

$$\hat{f}(i) = \frac{1}{\pi_i}\langle q_i, f\sum_j q_j\rangle = \int \nu_i(x) f(x) dx = \mathbb{E}_\mu\left[f(X_t)\middle|\tilde{X}_t^- = i\right] \tag{45}$$

where $\nu_i(x) = \frac{q_i(x)\mu(x)}{\pi_i} = \mathbb{P}(X_t = x|\tilde{X}_t^- = i)$ is the probability density of finding the system in state x given that it came last from i. Hence $\hat{f}(i)$ is the average costs conditioned on the information $\tilde{X}_t^- = i$, i.e. X_t came last from A_i, which is the natural extension to the full partition case where $\hat{f}(i)$ was the average costs conditioned on the information that $X_t \in A_i$.

The matrix $K = \pi_i^{-1}\langle q_i, Lq_j\rangle$ is reversible with stationary distribution

$$\pi_i = \langle q_i, \mathbb{1}\rangle = \mathbb{P}_\mu(\tilde{X}_t^- = i)$$

and is related to core MSMs again:

$$K = \lim_{\tau\to 0}\frac{1}{\tau}\left(P^\tau - M\right)$$

where P^τ and M are now the matrices for core MSMs as in (18). Formally, K is the generator of the P^τ, but these do not form a semigroup since $M \neq \mathbb{1}$, and therefore we cannot interpret K directly as e.g. the generator of $\tilde{X}_t^-$. Nevertheless, the entries of K are the transition rates between the core sets as defined in transition path theory [42]. We can sample P^τ and M using (20) and (21), and because we used an incomplete partition, the recrossing problem is removed, and there is no difficulty in sampling P^τ for all lagtimes τ and therefore K directly. It is worth noting that F can also be sampled:

$$F_{ij} = \mathbb{E}_\mu\left[f(X_t)\chi_{\{\tilde{X}_t^+ = j\}}\middle|\tilde{X}_t^- = i\right]$$

Therefore, as in the construction of core MSMs, we do not need to compute committor functions explicitly. Note however that $G \neq L$, there is a reweighting due to the overlap of the q_i's which causes F to be nondiagonal. This reweighting is the surprising bit of this discretization. From properties (M1)-(M3) from section 7 we see however that G and K are both reversible with stationary distribution π. Finally, note that if the cost function $f(x)$ doesn't satisfy $\|f\|_\infty \leq C$ from (M3), G will not even be a generator matrix. In this case (34) still has a solution $\hat{\phi}$ which is the bestapproximation to ϕ, but this solution may not be unique, it may not satisfy $\hat{\phi} > 0$, and we have no interpretion as a discrete control problem.

9 Numerical Results

9.1 1D Potential Revisited

Firstly, we study diffusion in the triple well potential which is presented in Figure 2. This potential has three mimima at approximately $x_{0/1} = \pm 3.4$ and $x_2 = 0$. We choose the three core sets $C_i = [x_i - \delta, x_i + \delta]$ around the minima with $\delta = 0.2$. Take τ to be the first hitting time of C_0. We are interested in the moment generating function $\phi(x) = \mathbb{E}\left[e^{-\epsilon^{-1}\sigma\tau}\right]$ of passages into C_0 and the cumulant generating function $W = \epsilon \log \phi$. This is of the form (32) for $A = C_i$ and $f = \sigma$ a constant function.

In Figure 10a the potential V and effective potential U are shown for $\beta = 2$ and $\sigma = 0.08$ (solid lines), cf. equation (30). One can observe that the optimal control effectively lifts the second and third well up which means that the optimal control will drive the system into C_0 very quickly. The reference computations here have been carried out using a full partition FEM discretization of (31) with a lattice spacing of $h = 0.01$. Now we study the MJP approximation constructed via the committor functions shown in Figure 10b. These span a three-dimensional subspace, but due to the boundary conditions the subspace D of the method is actually two-dimensional. The dashed line in Figure 10a gives the approximation to U calculated by solving (38). We can observe extremely good approximation quality, even in the transition region. In Figure 10c the approximation to the optimal control $\alpha^*(x)$ (solid line) and its approximation $\hat{\alpha}^* = -\sqrt{2}\nabla\hat{W}$ (dashed line) are shown. The core sets are shown in blue. We can observe jumps in $\hat{\alpha}^*$ at the left boundaries of the core sets. This is to be expected and comes from the fact that the committor functions are not smooth at the boundaries of the core sets, but only continuous. Therefore the approximation to U is continuous, but the approximation to α^* is not.

Next we construct a core MSM to sample the matrices K and F. 100 trajectories of length $T = 20000$ were used to build the MSM. In Figure 10d, W and its estimate using the core MSM is shown for $\epsilon = 0.5$ and different values of σ. Each of the 100 trajectories has seen about four transitions. For comparison, a direct sampling estimate of W using the same data is shown (green). The direct sampling estimate suffers from a large bias and variance and is practically useless. In contrast, the MSM estimator for W performs well for all considered values of σ and always its variance is significantly small. The constant C which ensures $\hat{\phi} > 0$ when $\sigma \leq C$ is approximately 0.2 in this case. This seems restrictive but still allows to capture all interesting information about ϕ and W.

9.2 Alanine Dipeptide

Lastly, we study α-β-transitions in Alanine dipeptide, a well-studied test system for Molecular Dynamics applications. We use a $1\mu s$ long trajectory simulated with the CHARMM 27 force field. The conformational dynamics is monitored as usual via the backbone dihedral angles ϕ and ψ. The data was first presented in [27]. We construct a full partition MSM with 250 clusters using k-means clustering. We are interested in the MFPT $\hat{t}(i) = \mathbb{E}_i[\tau_\alpha]$ where τ_α is the first hitting time of the α conformation, which we define as a circle with radius $r = 45$ around $(\phi_\alpha, \psi_\alpha) = (-80, -60)$. The MFPT vector $\hat{t}$ solves the boundary value problem

$$K\hat{t} = -1 \text{ outside of } \alpha, \quad \hat{t} = 0 \text{ in } \alpha,$$

but since K is not available directly via sampling, we have to consider the equation

$$\frac{1}{\tau}\left(P^\tau - 1\right)\hat{t} = -1 \text{ outside of } \alpha, \quad \hat{t} = 0 \text{ in } \alpha$$

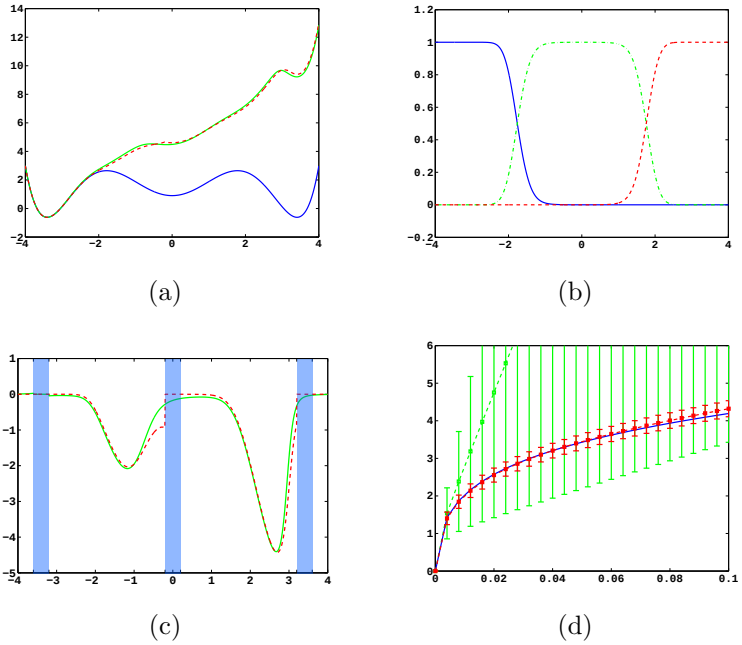
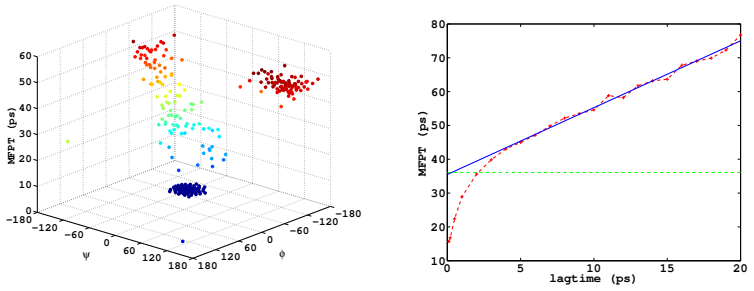


Figure 10: Three well potential example for $\epsilon = 0.5$ and $\sigma = 0.08$. (a) Potential $V(x)$ (blue), effective potential $U = V + 2W$ (green) and approximation of U with committors (dashed red). (b) The three committors $q_1(x)$, $q_2(x)$ and $q_3(x)$. (c) The optimal control $\alpha^*(x)$ (solid line) and its approximation (dashed line). Core sets are shown in blue. (d) Optimal cost W for $\beta = 2$ as a function of σ. Blue: Exact solution. Red: Core MSM estimate. Green: Direct sampling estimate.

instead. The result will depend on the choice of lagtime τ. In Figure 11a, the results are shown for $\tau = 5$, we can identify the β-structure as the red cloud of clusters where $\hat{t}(i)$ is approximately constant. In Figure 11b, $\hat{t}_{\beta\alpha} = \mathbb{E}(\hat{t}(i)|i \in \beta)$ is shown as a function of τ. We observe a linear behavior for large τ which is due to the linear error introduced in the replacement of K with $\frac{1}{\tau}(P^\tau - 1)$ and a nonlinear drop for small τ which is due to Non-Markovianity. Our best guess is therefore a linear interpolation to $\tau = 0$, which is indicated by the solid line. The result is $\hat{t}^{(0)}_{\beta\alpha} = 35.5ps$. As a comparison the reference value $\hat{t}^{ref}_{\beta\alpha} = 36.1ps$ from [27] is shown as a dashed line. It was computed in [27] as an inverse rate, using the slowest ITS and information about the equilibrium weights of the α and *beta* structure. We see very good agreement. The result is of course dependent though on the assignment of clusters to the α and β structure. Some tests show that $\hat{t}^{(0)}_{\beta\alpha}$ as computed with the interpolation method is fairly insensitive to this choice.

In [14] it is demonstrated how to use the method presented herein for maximizing the population of the α-conformation of Alanin dipeptide based on the MSM used here.



(a) (b)

Figure 11: Dipeptide example. (a) MFPT from β to α in ϕ-ψ space for $\tau = 5$. The red cloud to the right is the β-structure. (b) MFPT as a function of τ (dashed line) and linear interpolation to $\tau = 0$ (solid line). Green dashed line: Reference computed via slowest ITS.

10 Conclusion

In this article, we have discussed an approach to overcome direct sampling issues of rare events in molecular dynamics based on spatial discretization of the molecular state space. The strategy is to define a discretization by subsets of state space such that the samling effort with respect to transitions between the sets is much lower than the direct estimation of the rare events under consideration. That is, without having to simulate rare events we construct a so called Markov State Model, a Markov chain approximation to the original dynamics. Since the state space of the MSM is finite, we can then calculate the properties of interest by simply solving linear systems of equations. Of course, it is crucial that these properties of the MSM can be related to the rare event properties of the original process that we have not been able to sample directly.

This is why we have analyzed the approximation quality of MSMs in the first part of the article. We have used the interpretation of MSMs as projections of the transfer operator to (1) derive conditions that guarantee an accurate reproduction of the dynamics, and (2) show how to construct models based on a core set discretization by leaving the state space partly undiscretized.

In the second part of the article, we have used the concept of MSM discretization to solve MD optimal control problems in which one computes the optimal external force that drives the molecular system to show an optimized behavior (maximal possible population in a conformation; minimal mean first passage time to a certain conformation) under certain constraints. We have demonstrated that the spatial discretization underlying an MSM turns the high-dimensional continuous optimal control problem into a rather low-dimensional discrete optimal control problem of the same form that can be solved efficiently. This result allows two different types of application: (1) If one can construct an MSM for a molecular system in equilibrium, then one can use it to compute optimal controls that extremize a given costs criterion. (2) If an MSM can be computed based on transition probabilities between neighboring core sets alone then the rare event statistics for transitions between strongly separated metastable states of the system can be computed from an associated optimal control problem that can be solved after discretization using the pre-computed MSM.

References

[1] C. Hartmann, R. Banisch, M. Sarich, T. Badowski, and C. Schütte. Characterization of rare events in molecular dynamics. *Entropy*, 2013. submitted.

[2] P. Metzner, Ch. Schütte, and E. Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219, 2009.

[3] Albert C. Pan and Benoît Roux. Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.*, 129(6):064107+, August 2008.

[4] F. Noé, Ch. Schütte, E. Vanden-Eijnden, L. Reich, and T. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium trajectories. *PNAS*, 106(45):19011–19016, 2009.

[5] Ch. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Applications to Biomolecules*. Habilitation thesis, Fachbereich Mathematik und Informatik, FU Berlin, 1998.

[6] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A Direct Approach to Conformational Dynamics based on Hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[7] J.H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, C. Schuette, John D. Chodera, and Frank Noe. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 2010. submitted.

[8] F. Noé, I. Horenko, C. Schütte, and J. C. Smith. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.*, 126:155102, 2007.

[9] J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126:155101, 2007.

[10] Nicaolae V. Buchete and Gerhard Hummer. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B*, 112:6057–6069, 2008.

[11] Jan-Hendrik Prinz, Bettina Keller, and Frank Noé. Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.*, 13:16912–16927, 2011.

[12] Bettina Keller, Jan-Hendrik Prinz, and Frank Noé. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chem. Phys. (in press)*, 2011. Accepted 31. August 2011.

[13] G. Bowman, V. Volez, and V. S. Pande. Taming the complexity of protein folding. *Current Opinion in Structural Biology*, 21:4–11, 2011.

[14] Ch. Schütte, S. Winkelmann, and C. Hartmann. Optimal control of molecular dynamics using markov state models. *Math. Program. Series B*, 134:259–282, 2012.

[15] C. Hartmann and C. Schütte. Efficient rare event simulation by optimal nonequilibrium forcing. *J. Stat. Mech. Theor. Exp.*, page 11004, 2012.

[16] M. Jäger, Y. Zhang, J. Bieschke, H. Nguyen, M. Dendle, M. E. Bowman, J. P. Noel, M. Gruebele, and J. W. Kelly. Structure-function-folding relationship in a ww domain. *Proc. Natl. Acad. Sci. USA*, 103:10648–10653, 2006.

[17] Andrei Y. Kobitski, Alexander Nierth, Mark Helm, Andres Jäschke, and G. Ulrich Nienhaus. Mg2+ dependent folding of a Diels-Alderase ribozyme probed by single-molecule FRET analysis. *Nucleic Acids Res.*, 35:2047–2059, 2007.

[18] S. Fischer, B. Windshuegel, 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*, 102:6873–6878, 2005.

[19] F. Noé, D. Krachtus, J. C. Smith, and S. Fischer. Transition networks for the comprehensive characterization of complex conformational change in proteins. *J. Chem. Theo. Comp.*, 2:840–857, 2006.

[20] Andreas Ostermann, Robert Waschipky, Fritz G. Parak, and Ulrich G. Nienhaus. Ligand binding and conformational motions in myoglobin. *Nature*, 404:205–208, 2000.

[21] W. Huisinga. *Metastability of Markovian Systems A transfer operator based approach in application to molecular dynamics.* Phd thesis, Fachbereich Mathematik und Informatik, FU Berlin, 2001.

[22] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6:399–424, 2004.

[23] M. G. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J. Reine Angew. Math.*, 134:198–287.

[24] M. Sarich, F. Noé, and Ch. Schütte. On the approximation quality of Markov state models. *Multiscale Modeling and Simulation*, 8(4):1154–1177, 2010.

[25] Marco Sarich. *Projected Transfer Operators.* PhD thesis, Freie Universitt Berlin, 2011.

[26] M. Sarich and Ch. Schütte. Approximating selected non-dominant timescales by Markov state models. *Comm. Math. Sci.*, 10, 2012.

[27] Ch. Schütte, F. Noé, Jianfeng Lu, M. Sarich, and E. Vanden-Eijnden. Markov state models based on milestoning. *J. Chem. Phys*, 134 (19), 2011.

[28] Anton K. Faradjian and Ron Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120:10880–10889, 2004.

[29] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Fischbach, M. Held, J. D. Chodera, Ch. Schtte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134:174105, 2011.

[30] Susanna Roeblitz. *Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics.* PhD thesis, FU Berlin, 2008.

[31] N. Djurdjevac, M. Sarich, and Ch. Schütte. On Markov state models for metastable processes. *Proceeding of the ICM 2010 as invited lecture*, 2010. Download via http://www.math.fu-berlin.de/groups/biocomputing/publications/index.html.

[32] I. Horenko, E. Dittmer, F. Lankas, J. Maddocks, Ph. Metzner, and Ch. Schütte. Macroscopic dynamics of complex metastable systems: Theory, algorithms, and application to B-DNA. *J. Appl. Dyn. Syst.*, 2009.

[33] W.H. Fleming and H.M. Soner. *Controlled Markov Processes and Viscosity Solutions.* Springer; 2nd edition, 2005.

[34] B. Oksendal. *Stochastic Differential Equations.* Springer, 2003.

[35] Paolo Pra, Lorenzo Meneghini, and WolfgangJ. Runggaldier. Connections between stochastic control and dynamic games. *Mathematics of Control, Signals and Systems*, 9:303–326, 1996.

[36] H.J. Kushner and P.G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer Verlag, 1992.

[37] M. Braack. *Finite Elemente*. 2012.

[38] R. Banisch and C. Hartmann. A meshfree discretization for optimal control problems. *to appear*, 2013.

[39] S. Sheu. Stochastic control and exit probabilities of jump processes. *SIAM Journal on Control and Optimization*, 23(2):306–328, 1985.

[40] J.C. Latorre, Ph. Metzner, C. Hartmann, and Ch. Schtte. A structure-preserving numerical discretization of reversible diffusions. *Commun. Math. Sci.*, 9:1051–1072, 2011.

[41] J. D. Chodera, P. J. Elms, W. C. Swope, J.-H. Prinz, S. Marqusee, C. Bustamante, F. Noé, and V. S. Pande. A robust approach to estimating rates from time-correlation functions. *arXiv: 1108.2304*, August 2011.

[42] E. Vanden-Eijnden. Transition path theory. *Lect. Notes Phys.*, 703:439–478, 2006.