

Article

## Markov State Models for Rare Events in Molecular Dynamics

Marco Sarich <sup>1,\*</sup>, Ralf Banisch <sup>1</sup>, Carsten Hartmann <sup>1</sup> and Christof Schütte <sup>1,2</sup>

<sup>1</sup> Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, Berlin 14195, Germany; E-Mails: ralf.banisch@fu-berlin.de (R.B.); chartman@mi.fu-berlin.de (C.H.); christof.schuette@fu-berlin.de (C.S.)

<sup>2</sup> Zuse Institute Berlin, Takustr. 7, Berlin 14195, Germany

\* Author to whom correspondence should be addressed; E-Mail: sarich@math.fu-berlin.de; Tel.: +49 30 838 75 322.

Received: 18 September 2013; in revised form: 3 December 2013 / Accepted: 9 December 2013 / Published: 30 December 2013

---

**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

---

### 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 in which is interested. 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 Modeling, *avoids* the combinatorial explosion of the number of discretization elements with the increasing size of the molecular system in contrast to other methods for spatial discretization.

The actual construction of an MSM requires one to sample certain transition probabilities of the underlying dynamics between sets. The idea is: (1) to choose the sets such that the sampling 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 original 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 Modeling for realistic molecular systems, the transition probabilities and the respective statistical uncertainties are estimated from short molecular dynamics (MD) trajectories only, *cf.* [7]. This makes Markov State Modeling 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 demonstrate 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.,  $E \subset \mathbb{R}^d$ . That is,  $X_t$  is the state of the molecular system at time  $t$  resulting from any usually used form 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 Modeling is to derive a Markov chain,  $(\hat{X}_k)_{k \in \mathbb{N}}$ , on a finite and preferably small state space  $\hat{E} = \{1, \dots, n\}$  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,  $(\hat{X}_k)_{k \in \mathbb{N}}$ , lives on a finite state space, the construction of an MSM boils down to the computation of its transition matrix,  $P$ :

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

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 the 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 an 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 the 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  $A_1, \dots, A_n \subset E$  in continuous state space that form a full partition, *i.e.*:

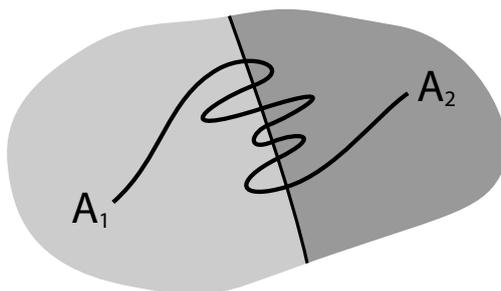
$$A_i \cap A_j = \emptyset \text{ for } i \neq j, \quad \bigcup_{i=1}^n A_i = E \quad (2)$$

Typical choices for such sets are box discretizations or Voronoi tessellations [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. This 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,  $P^\tau$ :

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

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  $\tau$ -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,  $\mu$ , and that it is time-reversible. Then, for any time-step,  $t \geq 0$ , we define the transfer operator,  $T_t$ , via the property:

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

as an operator  $T_t : L^2(\mu) \rightarrow L^2(\mu)$ . 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 other than the propagator of densities under the dynamics, but the densities are understood as densities with respect to the measure,  $\mu$ . 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:

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

The benefit of working with  $\mu$ -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 Equation (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  $\chi_1, \dots, \chi_n$  is a basis of  $D$ . Then, the so-called projected transfer operator,  $QT_\tau Q : D \rightarrow D$ , 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  $i = 1, \dots, n$ , 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 Equation (4). The subscript,  $\mu$ , shall indicate that  $X_0 \sim \mu$ . Therefore, 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 that are constant on the partitioning sets. This interpretation of an MSM is useful, since it allows one to analyze its approximation quality. For example, in [25,26], it is proven that we can reproduce an eigenvalue,  $\lambda$ , 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,  $\lambda_i$ , of the transfer operator directly relates to an implied timescale,  $\mathcal{T}_i$ , of the system via:

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

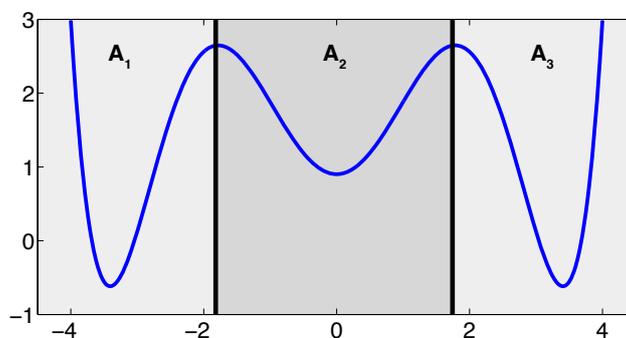
Therefore, the transition matrix Equation (4) that we construct from transitions between the sets,  $A_1, \dots, A_n$ , will generate a Markov chain that will reproduce the original timescales well if the partitioning sets are chosen such that the corresponding eigenvectors are almost constant on these 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,  $\delta$ , 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  $\varepsilon > 0$ .

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



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

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

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

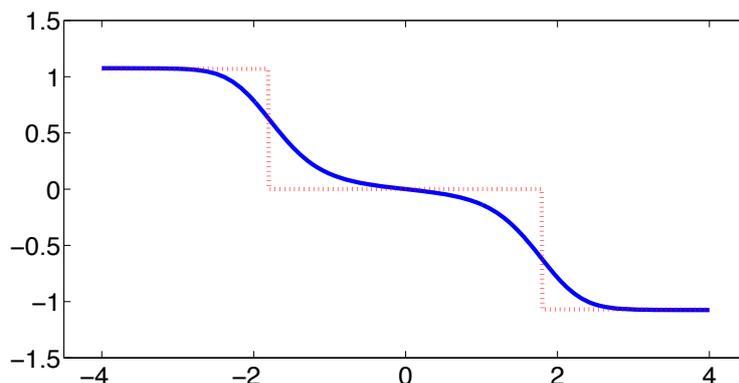
As one can see, the timescales are strongly underestimated. This is a typical phenomenon. From a statistical point of view, the recrossing problem will lead to cumulatively appearing transition counts when one computes the transition probabilities,  $\mathbb{P}_\mu[X_\tau \in A_j | X_0 \in A_i]$ , from a trajectory  $(X_t)$ , as discussed above. Therefore, on average, transitions between sets seem to become too likely, and hence,

**Table 1.** Comparison of implied timescales

	$T_1$	$T_2$
Original	103.7608	11.9566
Full partition 3 sets	80.6548	9.8784

the processes in the coarse-grained system get accelerated. We have seen in Equation (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.

**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  $A_1, A_2, A_3$ .



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. Therefore, we have two explanations of 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

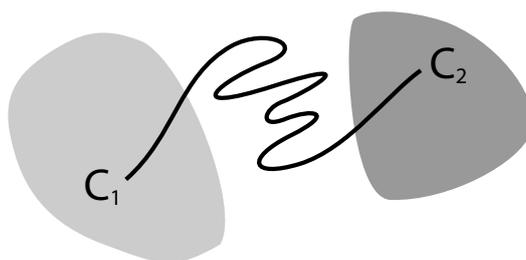
From Equation (10), we know how to compute a matrix representation for a projected transfer operator for an arbitrary subspace,  $D \subset L^2(\mu)$ . For a given basis,  $\chi_1, \dots, \chi_n$ , we have to compute Equations (11) and (12), so:

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

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 set Equation (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 other 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 Equation (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*,  $C_1, \dots, C_n \subset E$ , 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,  $\chi_1, \dots, \chi_n$ . Assume  $T_\tau$  is, again, a self-adjoint transfer operator and consider  $n$  core sets  $C_1, \dots, C_n$ . For every  $i$ , take the committor function,  $\chi_i$ , 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 projection,  $Q$ , onto the space spanned by these committor functions, the two following properties hold [25,27].

(P1) The matrices,  $M$  and  $P$ , in Equation (10) can be written as:

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

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  $t = k\tau$ .

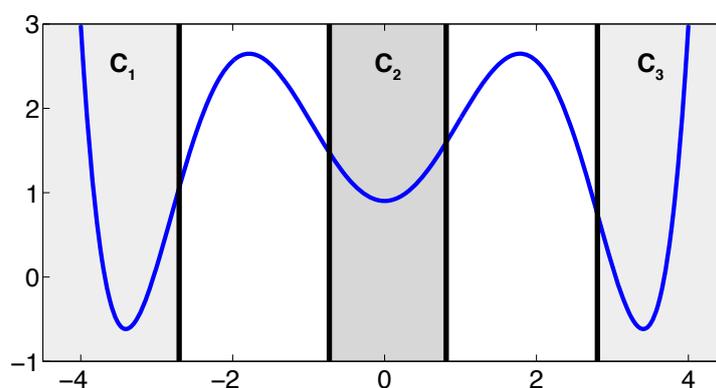
(P2) Let  $u_i$  be an eigenvector of  $T_\tau$  that is almost constant on the core sets. Let the region  $C = E \setminus \bigcup_i C_i$  that is not assigned to a core set be left quickly enough, so  $\mathbb{E}_x[\tau(C^c)] \ll \mathcal{T}_i$  for all  $x \in C$ , where  $\mathcal{T}_i$  is the timescale 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 taking 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 discretization. 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.

**Figure 5.** Excluding a small region of state space from the sets,  $A_1, A_2, A_3$ , as in Figure 2, to form core sets  $C_1, C_2, C_3$  that do not share boundaries anymore.



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:

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

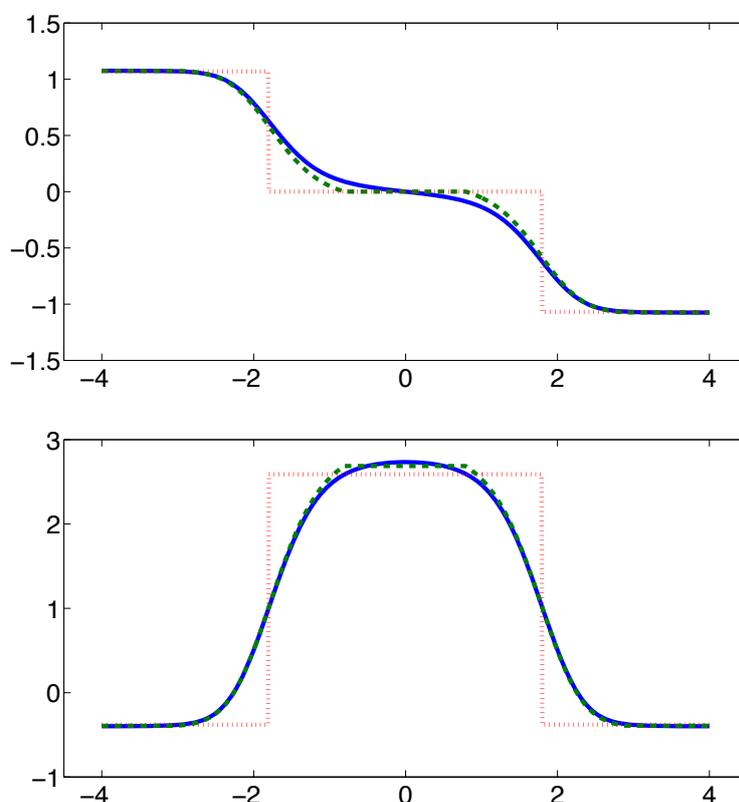
Comparing to the MSM for the full partition one can see that transitions between indices  $i$  and  $j$ ,  $i \neq j$  are less likely. Table 2 shows this leads to a far more accurate reproduction of the timescales in the system.

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  $T_2$ . Therefore, the committor functions should deliver a good approximation of the first two eigenvectors. Figure 6 underlines this theoretical result.

**Table 2.** More accurate approximation if implied timescales

	$T_1$	$T_2$
Original	103.7608	11.9566
3 core sets	100.8066	11.9145
Full partition 3 sets	80.6548	9.8784

**Figure 6. (Upper panel)** The first non-trivial eigenvector,  $u_1$  (solid blue), and its projection,  $Q_f u_1$  (finely dashed red), onto step functions (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,  $u_2$ .



## 5. Practical Considerations and MD Applications

In the previous sections, we have interpreted the construction of an MSM as a 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  $x_0, x_1, \dots, x_N$  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,  $A_1, \dots, A_m$ , and a lag time  $\tau = nh$  the discrete trajectory  $s_k = i \Leftrightarrow x_k \in A_i$  and compute the matrix,  $\hat{P}$ :

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

It is well known [7] that  $\hat{P}$  is a maximum likelihood estimator for the full partition MSM transition matrix Equation (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  $C_1, \dots, C_m$ , a lag time  $\tau = nh$  as before, and we define discrete milestoning trajectories by:

$$s_k^- = i \Leftrightarrow x_k \in A_i \text{ or came last from } A_i \text{ before time } k$$

$$s_k^+ = i \Leftrightarrow x_k \in A_i \text{ or went next to } A_i \text{ after time } k$$

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

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

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

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 [7,29].

Besides the choice of discretization and the available statistics, the estimates above also depend on a lag time,  $\tau$ . This dependence can be used to validate an MSM by a Chapman–Kolmogorov test [7]. 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 one 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:

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

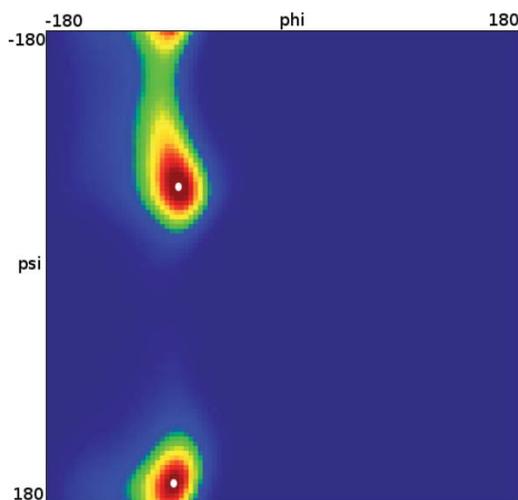
where  $N_{ij}^T$  is the amount of time in  $[0, T]$  the process has spent on its way from core set  $C_i$  to  $C_j$  and  $R_i^T$  is the total time in  $[0, T]$  the process came last from  $C_i$ . On the other hand, one can see [27,30] that  $K^* = KM^{-1}$  with the mass matrix,  $M$ , from above Equation (18), can be interpreted as a projection of the original generator of the process and, also, as a derivative of the core set MSM from above, *i.e.*:

$$K^* = \lim_{\tau \rightarrow 0} \frac{PM^{-1} - I}{\tau} \quad (23)$$

where  $P$  depends on  $\tau$  Equation (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  $\phi$  and  $\psi$  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,  $x_\alpha, x_\beta$ , 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,  $x_\alpha, x_\beta$ , in the angular space as white dots.

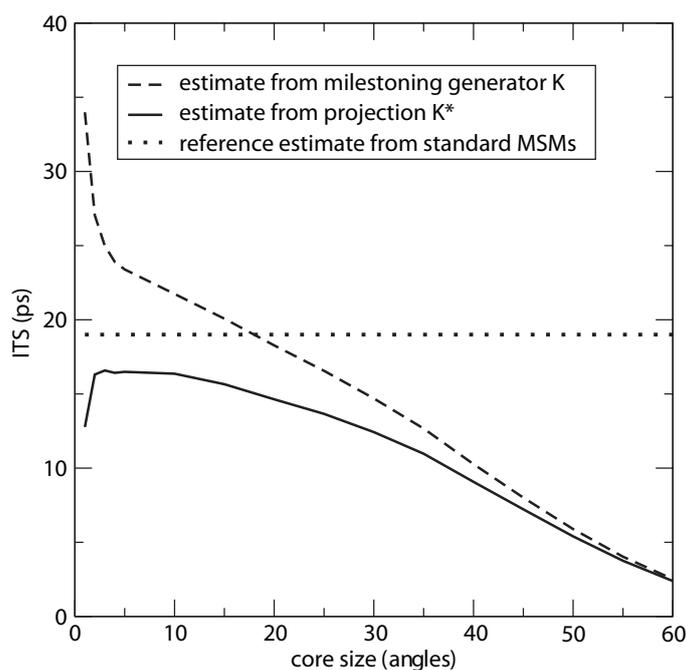


For computing a reference timescale, several MSMs based on three different full partitions using 10, 15 and 250 sets have been constructed for increasing lag times. In [27], it is shown that in each setting, the estimate for the longest implied timescale of the process converged to  $\approx 19$  ps for large enough  $\tau$ . Now, the implied timescales for the two different generators,  $K$  Equation (22) and  $K^*$  Equation (23), are computed. In Figure 8, the resulting timescales are plotted against the reference timescale  $\approx 19$  ps for varying size of the core sets.

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 generator,  $K^*$ , 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 as 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 milestone 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.

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



## 6. Further Applications in MD

Markov State Modeling has been shown to apply successfully to many different molecular systems, like peptides, including time-resolved spectroscopic experiments [10–12], proteins and protein folding [4,9,13], DNA [31] and ligand-receptor interaction [32]. In most of the respective publications, full partition MSMs are used, and the underlying discretization is based on cluster finding methods (see [7] for a review), while the sampling issues are tackled by means of ideas from enhanced sampling [33] and based on ensembles of rather short trajectories instead of one long one, *cf.* [4]. Core set-based approaches have been used just recently [10,27]; related algorithms are less well developed. However, recent work has shown that and how every full partition MSM can be easily transformed into a core set-based MSM with significantly *improved* approximation quality [34], making core set MSMs the most promising next generation MSM tools.

**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,  $\tau$ , 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:

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

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:

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

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 the “first order” transition probability of a system correctly enough and ignore all “higher order” ones, then our accuracy will not be spoiled. This nicely illustrates the main advantage of MSM modeling compared to classical long-term simulation: since only neighboring core sets have to be connected by accurately estimated rates, the long residence time of long-term trajectories between and in core sets can be avoided, thus cutting down total simulation time.

**Computation from Trajectories.** Clearly, constructing and analyzing a core set MSM will only have a computational advantage compared to the direct sampling of a rare event if the transition events between neighboring core sets occur on a much shorter timescale than the rare event itself. One should note that from the purely theoretical point of view, it would be optimal to choose only very few core sets in the most metastable regions of state space, because this would minimize the projection error  $\delta = \|u - Qu\|$  for each dominant eigenvector  $u$ , as discussed in Section 3. On the other hand, when estimating the MSM from trajectories, only a finite amount of statistics will be available, so there will also be a statistical error. In order to keep the total error small, additional core sets in less metastable parts of state space typically have to be introduced. In the end, this makes the estimation of a core set MSM possible without having to sample rare events. Note that the projection error is still under control, as long as there is a transition region between the core sets that is typically left very quickly (see Property (P2) in Section 4).

In practice, the statistics of the transition events between core sets will preferably be estimated from many short trajectories using milestone techniques [27,28] and parallel computing. However, any algorithm for the construction of a core set MSM has to find a balance between sampling issues (not too many too long trajectories needed) and discretization issues (not too many core sets). Construction of such an algorithm still is ongoing research.

This article cannot give a detailed review on the algorithmic realization of MSMs for realistic molecular systems and on the findings resulting from such applications, since this is discussed to some extent elsewhere; see [7] for a recent review of the algorithmic aspects and [32,35] for ligand-receptor interaction.

### 7. 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., the sampling of rare events. Specifically, we consider the case that  $(X_t)_{t \geq 0}$  is the solution of:

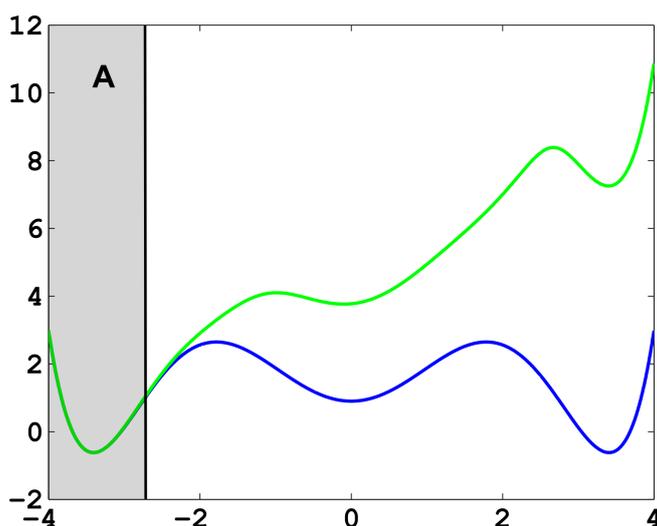
$$dX_t = (\sqrt{2}u_t - \nabla V(X_t))dt + \sqrt{2\varepsilon}dB_t \tag{26}$$

with potential  $V$ , Brownian motion  $B_t$  and temperature  $\varepsilon > 0$ , as in Equation (16), and an unknown control variable,  $u: [0, \infty) \rightarrow \mathbb{R}^d$ , 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  $\tau < \infty$  (almost surely) is a random stopping time that is determined by  $X_t$  hitting a given target set,  $A \subset E$ , i.e.,  $\tau = \inf\{t > 0: X_t \in A\}$ , in other words, we are interested in controlling  $X_t = X_t^u$  until it reaches  $A$ . As an example, consider the case  $f = 1$  with the potential considered in Figure 5 and the target region,  $A$ , around the left well. This situation is illustrated in Figure 9 and amounts to the situation that one seeks to minimize the time to reach  $A$  by tilting the potential towards  $A$ ; tilting the potential too much is prevented by the quadratic penalization term in the cost functional that grows when too much force is applied.

**Figure 9.** The potential from Figure 5 (blue) and a tilted potential to minimize the time required to hit the target set,  $A$  (green).



Other choices of  $f$  in Equation (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,  $\tau \rightarrow \infty$ , and the construction would follow analogously. For compactness, we consider here only cost functions as in Equation (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,  $\alpha: E \rightarrow \mathbb{R}^d$ . Moreover, only controls with finite energy are considered, for otherwise,  $J(u; x) = \infty$ . For control problems of the form Equations (26) and (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 the Hamilton–Jacobi–Bellman type (see [36]):

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

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 Equation (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  $\phi$  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 [37], the solution to Equation (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 Equation (26) can be computed by solving Equation (31), which can be done *in principle* via Monte Carlo approximation of the expected value in Equation (32) if critical slowing down by rare events can be avoided.

**Remark.** The optimization problem Equation (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:

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

where the notation “ $Q \ll P$ ” means that  $Q$  has a density (That is, the density function,  $dQ/dP$ , exists and is almost everywhere positive and normalized) 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 Equation (26); in this sense, the notation in Equation (33) is meaningful. The second term:

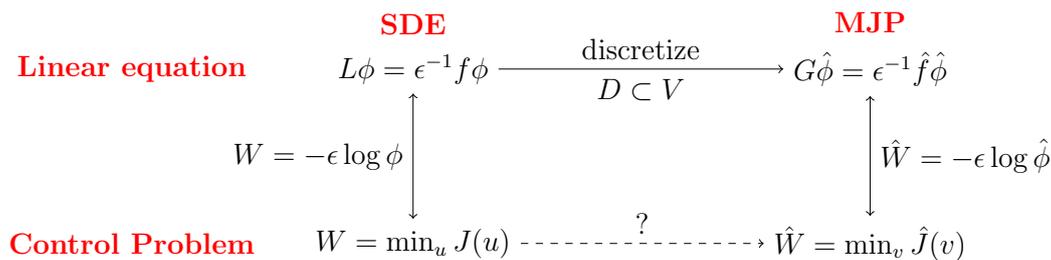
$$H(Q||P) = \varepsilon \int \log \left( \frac{dQ}{dP} \right) dQ$$

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 [38] or the article [1] in this special issue.

### 8. MSM Discretization of Optimal Control Problems

The basic idea is now to choose a subspace,  $D \subset L^2(\mu)$ , with basis  $\chi_1, \dots, \chi_n$  as in Markov State Modeling 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 [39]. 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 Equation (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  $\chi_1, \dots, \chi_{n+1}$  with the following properties:

- (S1) The  $\chi_i$  form a partition of unity, that is  $\sum_{i=1}^{n+1} \chi_i = \mathbb{1}$ .
- (S2) The  $\chi_i$  are adapted to the boundary conditions in Equation (31), that is  $\chi_{n+1}|_A = 1$  and  $\chi_i|_A = 0$  for  $i \in \{1, \dots, n\}$ .

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

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

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

$$\sum_{j=1}^{n+1} (K_{ij} - \varepsilon^{-1} F_{ij}) \hat{\phi}_j = 0, \quad i \in \{1, \dots, n\} \tag{34}$$

$$\hat{\phi}_{n+1} = 1$$

that is the discrete analogue of Equation (31). The discrete solution  $\hat{\phi} = Q\phi$  is optimal in the sense of being the best approximation of  $\phi$  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 [40] for details. (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,  $\|\phi\|_A^2 = \varepsilon^{-1} \langle \phi, f \phi \rangle + \varepsilon \langle \nabla \phi, \nabla \phi \rangle$ .) 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 Equation (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 [41]. The bottom line of Equation (35) shows that discretizing Equation (31) via Equation (34) minimizes the projection error measured in the energy norm. Since all functions are  $\mu$ -weighted, the approximation will be good in regions visited with high probability and less good in regions with lower probability. The error estimate Equation (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  $\phi$ , provided that  $A$  leaves the subspace,  $D$ , almost invariant. As was pointed out earlier, this is exactly the case when the  $\chi_i$  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 Equation (36), will vanish if the  $\chi_i$  form an arbitrarily fine full partition of  $E$ . If we follow the core set idea from Section 4 and choose the  $\chi_i$  to be committor functions, we have good control over  $\|Q^\perp\phi\|_\mu$ . Due to [41]:

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

where  $C = E \setminus \cup_i C_i$  is the transition region,  $\kappa = \sup_{x \in C} \mathbb{E}_x \tau_{E \setminus C}$  is the maximum expected time of hitting the metastable set from outside (which is short) and  $P$  is the orthogonal projection onto the subspace  $V = \{v \in L^2(\mu), v = \text{const on every } C_i\} \subset L^2(\mu)$ . Note that  $P^\perp\phi = 0$  on  $C$ . The errors,  $\|P^\perp\phi\|_\mu$  and  $\|P^\perp\phi\|_\infty$ , measure how constant the solution,  $\phi$ , is on the core sets. Hence, Equation (37) together with Equation (36) gives us complete control over the approximation error of our projection method, even if we consider just a few core sets. In Section 9, we will investigate the full and core set partition cases further.

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

$$\sum_{j=1}^{n+1} (G_{ij} - \varepsilon^{-1}\Lambda_{ij}) \hat{\phi}_j = 0, \quad i \in \{1, \dots, n\} \tag{38}$$

$$\hat{\phi}_{n+1} = 1$$

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,  $\chi_i$ , form a partition of unity, i.e.,  $\sum_i \chi_i = \mathbb{1}$ . This implies that  $\pi$  is a probability distribution on the discrete state space  $\hat{E} = \{1, \dots, n + 1\}$ . We summarize properties of the matrices,  $K$ ,  $F$  and  $G$ ; see also [41]:

(M1)  $K$  is a generator matrix of an 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,  $\pi$ , 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 a row sum of 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 an MJP that we shall denote by  $(\hat{X}_t)_{t \geq 0}$ , and Equation (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 [42] that  $\hat{W}$  can be interpreted as the value function of a Markov decision problem with cost functional (cf. also [36]):

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

that is minimized over the set of Markovian control strategies,  $v: \hat{E} \rightarrow (0, \infty)$ , 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  $\tau = \inf\{t > 0: \hat{X}_t = n + 1\}$  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, Continued.** From [42], we know that the optimal cost:

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

is given by  $\hat{W} = -\varepsilon \log \hat{\phi}$ , where  $\hat{\phi}$  solves Equation (38), with the optimal feedback strategy given by  $v^*(i) = \hat{\phi}_i$  (see [36]). 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  $\pi^v$ , 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 Equation (33) in terms of the relative entropy:

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

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  $\mu$  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:

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

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  $E$ .

### 9. 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  $\chi_i$  satisfy Assumptions (S1) and (S2) discussed in Section 8. 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,  $A_i$ . Equation (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 neighbors ( $K_{ij} = 0$  otherwise). Here,  $m$  is the Lebesgue measure, and  $h_{ij}$ ,  $S_{ij}$  and  $\bar{x}_{ij}$  are defined as in Figure 10.  $K$  is the generator of an MJP on the cells,  $A_i$ , and coincides with the so-called *finite volume approximation* of  $L$  discussed in [43]. It is reversible with stationary distribution:

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

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

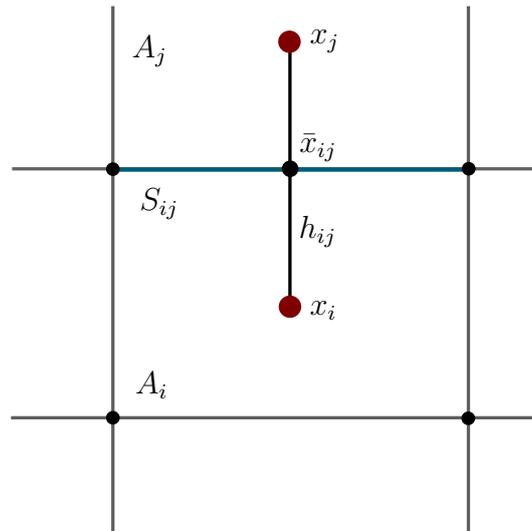
$$\lim_{\tau \rightarrow 0} \frac{1}{\tau} (P_{ij}^\tau - M_{ij}) = \lim_{\tau \rightarrow 0} \frac{1}{\pi_i} \langle \chi_i, \frac{1}{\tau} (T_\tau - \mathbf{1}) \chi_j \rangle = \frac{1}{\pi_i} \langle \chi_i, L \chi_j \rangle = K_{ij},$$

thus  $K$  is the generator of the semigroup of transition matrices,  $P^\tau$ . Therefore we could obtain  $K$  by sampling in the same way we obtained  $P^\tau$  through Equation (19) in Section 5. This is difficult, however, due to recrossing problems for small  $\tau$ ; see e.g., [44]. 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,

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

which establishes the connection to the continuous case. However, more is true: The whole discrete control problem reduces to first order in  $h$  to the well-known Markov chain approximation (MCA) [39], which allows us to use convergence theory for MCAs to conclude that, for  $n \rightarrow \infty$ , the optimal control and value function of the discrete control problem converge to their continuous counterparts. More details can be found in [41].

**Figure 10.** The mesh for the full partition.



**Core Set Partition.** Now, we choose core sets  $C_1, \dots, C_{n+1}$  with  $C_{n+1} = A$ , and we let  $\chi_i = q_i$  be the committor function of the process with respect to  $C_i$ , as in Section 4. These  $\chi_i$  satisfy Assumptions (S1) and (S2) discussed in Section 8. 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 Equation (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:

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

where  $P^\tau$  and  $M$  are now the matrices for core MSMs, as in Equation (18). Formally,  $K$  is the generator of the  $P^\tau$ , but these do not form a semigroup, since  $M \neq \mathbb{1}$ . 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 [45]. We can sample  $P^\tau$  and  $M$  using Equations (20) and (21); because we used an incomplete partition, the recrossing problem is removed, and there is no difficulty in sampling  $P^\tau$  for all lag times,  $\tau$ , and therefore,  $K$  directly. It is worth noting that  $F$  can also be sampled:

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

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 non-diagonal. This reweighting is the surprising bit of this discretization. From properties (M1)–(M3) from Section 8, we see, however, that  $G$  and  $K$  are both reversible with stationary distribution,  $\pi$ . Finally, note that if the cost function,  $f(x)$ , does not 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 best approximation to  $\phi$ , but this solution may not be unique; it may not satisfy  $\hat{\phi} > 0$ , and we have no interpretation as a discrete control problem.

## 10. Numerical Results

### 10.1. 1D Potential Revisited

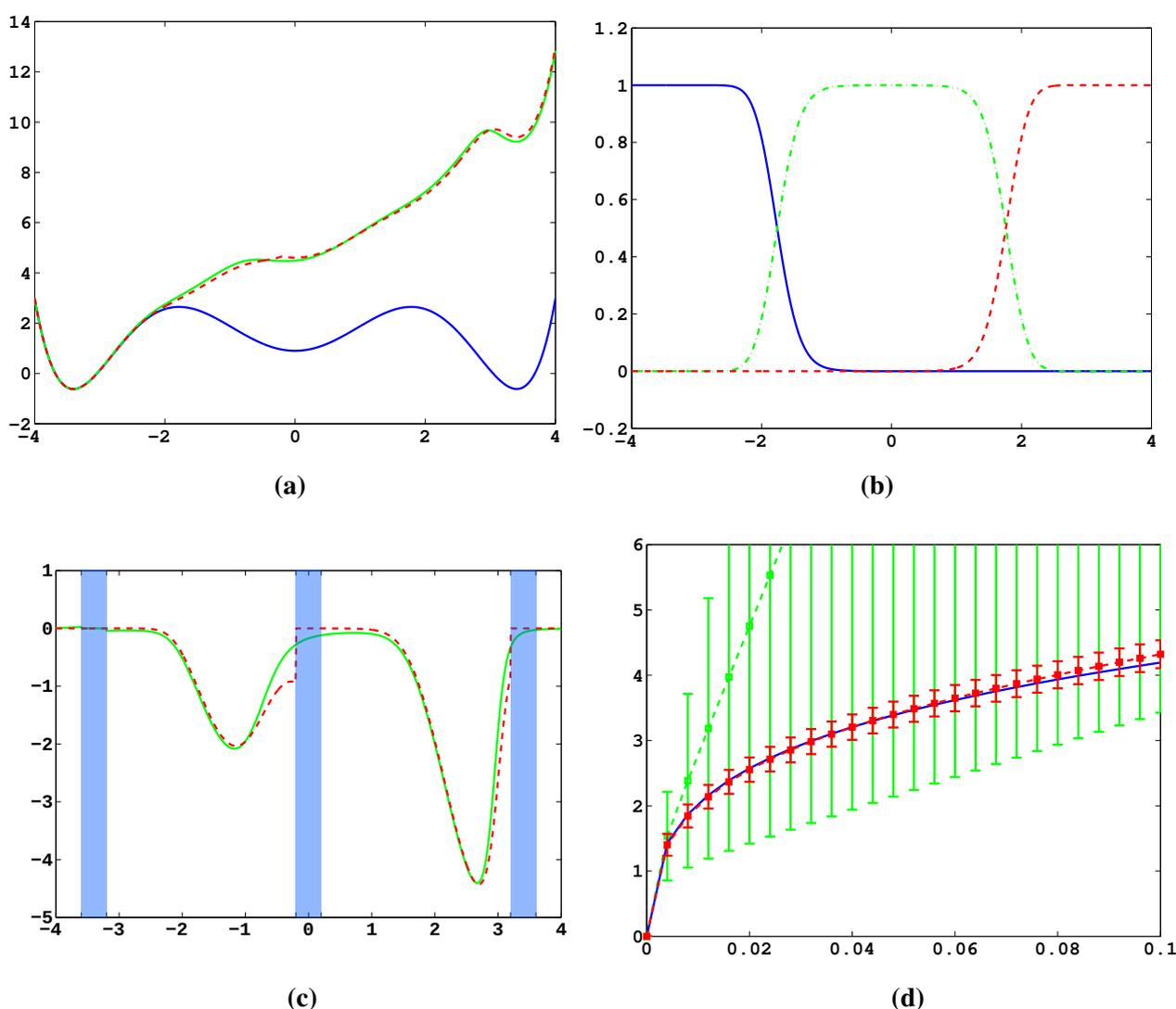
Firstly, we study diffusion in the triple well potential, which is presented in Figure 2. This potential has three minima at approximately  $x_{0/1} = \pm 3.4$  and  $x_2 = 0$ . We choose the three core sets  $C_i = [x_i - \delta, x_i + \delta]$  around the minima with  $\delta = 0.2$ . Take  $\tau$  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 Equation (32) for  $A = C_i$  and  $f = \sigma$ , a constant function.

In Figure 11a, 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 (finite element method) discretization of Equation (31) with a lattice spacing of  $h = 0.01$ . Now, we study the MJP approximation constructed via the committor functions shown in Figure 11b. 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 11a gives the approximation to  $U$  calculated by solving Equation (38). We can observe extremely good approximation quality, even in the transition region. In Figure 11c, 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  $\alpha^*$  is not.

Next, we construct a core MSM to sample the matrices,  $K$  and  $F$ . One hundred trajectories of length  $T = 20,000$  were used to build the MSM. In Figure 11d,  $W$  and its estimate using the core MSM are shown for  $\epsilon = 0.5$  and different values of  $\sigma$ . 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  $\sigma$ , 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 one to capture all interesting information about  $\phi$  and  $W$ .

**Figure 11.** 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  $\sigma$ . Blue: Exact solution. Red: Core MSM estimate. Green: Direct sampling estimate.



## 10.2. Alanine Dipeptide

Lastly, we study  $\alpha$ - $\beta$ -transitions in alanine dipeptide, a well-studied test system for molecular dynamics applications. We use a  $1\mu\text{s}$  long trajectory simulated with the CHARMM (Chemistry at HARvard Molecular Mechanics) 27 force field. The conformational dynamics is monitored as usual

via the backbone dihedral angles,  $\phi$  and  $\psi$ . 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 (mean first passage time)  $\hat{t}(i) = \mathbb{E}_i[\tau_\alpha]$ , where  $\tau_\alpha$  is the first hitting time of the  $\alpha$  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

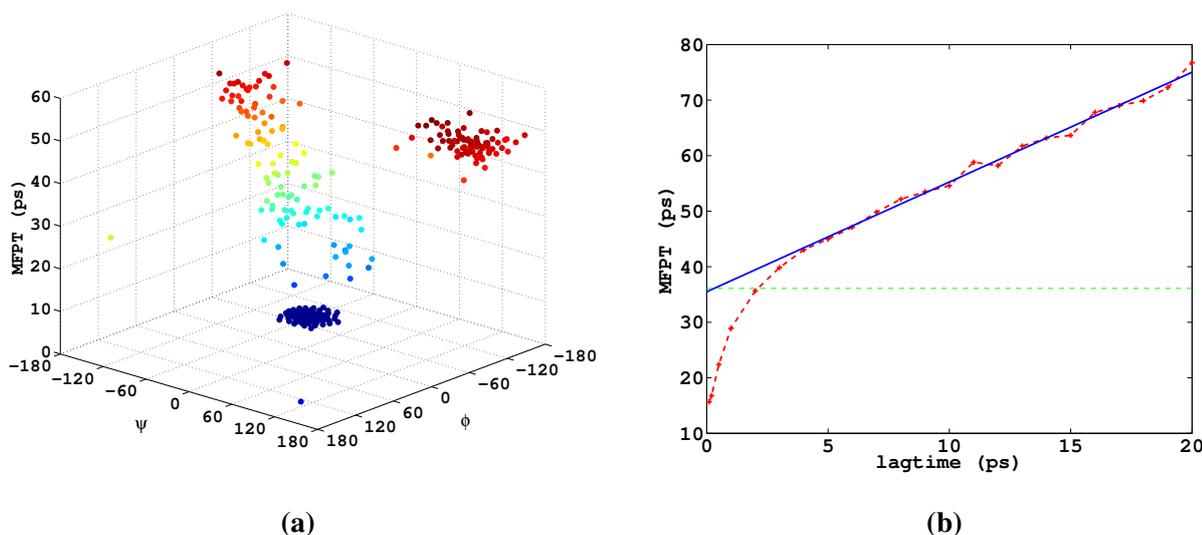
$$K\hat{t} = -1 \text{ outside of } \alpha, \quad \hat{t} = 0 \text{ in } \alpha$$

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

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

instead. The result will depend on the choice of lag time  $\tau$ . In Figure 12a, the results are shown for  $\tau = 5$ ; we can identify the  $\beta$ -structure as the red cloud of clusters where  $\hat{t}(i)$  is approximately constant. In Figure 12b,  $\hat{t}_{\beta\alpha} = \mathbb{E}(\hat{t}(i)|i \in \beta)$  is shown as a function of  $\tau$ . We observe a linear behavior for large  $\tau$ , 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  $\tau$ , 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  $\hat{t}_{\beta\alpha}^{(0)} = 35.5 \text{ ps}$ . As a comparison, the reference value  $\hat{t}_{\beta\alpha}^{ref} = 36.1 \text{ ps}$  from [27] is shown as a dashed line. It was computed in [27] as an inverse rate, using the slowest ITS (implied time scale) and information about the equilibrium weights of the  $\alpha$  and  $\beta$  structure. We see very good agreement. The result is, of course, dependent, though, on the assignment of clusters to the  $\alpha$  and  $\beta$  structure. Some tests show that  $\hat{t}_{\beta\alpha}^{(0)}$  as computed with the interpolation method is fairly insensitive to this choice.

**Figure 12.** Dipeptide example. (a) MFPT from  $\beta$  to  $\alpha$  in  $\phi$ - $\psi$  space for  $\tau = 5$ . The red cloud to the right is the  $\beta$ -structure. (b) MFPT as a function of  $\tau$  (dashed line) and linear interpolation to  $\tau = 0$  (solid line). Green dashed line: reference computed via the slowest ITS.



In [14], it is demonstrated how to use the method presented herein for maximizing the population of the  $\alpha$ -conformation of alanine dipeptide based on the MSM used here.

## 11. Conclusions

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 sampling 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 applications: (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.

## Acknowledgement

The authors have been supported by the DFG Research Center MATHEON.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Hartmann, C.; Banisch, R.; Sarich, M.; Badowski, T.; Schütte, C. Characterization of rare events in molecular dynamics. *Entropy* **2014**, *16*, 350–376.
2. Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Transition path theory for markov jump processes. *Multiscale Model. Simul.* **2009**, *7*, 1192–1219.

3. Pan, A.C.; Roux, B. Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.* **2008**, *129*, 064107.
4. Noé, F.; Schütte, C.; vanden-Eijnden, E.; Reich, L.; Weikl, T. Constructing the full ensemble of folding pathways from short off-equilibrium trajectories. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19011–19016.
5. Schütte, C. Conformational Dynamics: Modelling, Theory, Algorithm, and Applications to Biomolecules. Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, Berlin, Germany, 1998.
6. Schütte, C.; Fischer, A.; Huisinga, W.; Deuffhard, P. A direct approach to conformational dynamics based on hybrid monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146–168.
7. Prinz, J.H.; Wu, H.; Sarich, M.; Keller, B.; Fischbach, M.; Held, M.; Chodera, J.D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.
8. Noé, F.; Horenko, I.; Schütte, C.; Smith, J.C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **2007**, *126*, 155102.
9. Chodera, J.D.; Dill, K.A.; Singhal, N.; Pande, V.S.; Swope, W.C.; Pitera, J.W. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.
10. Buchete, N.V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.
11. Prinz, J.H.; Keller, B.; Noé, F. Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912–16927.
12. Keller, B.; Prinz, J.H.; Noé, F. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chem. Phys.* **2011**, in press.
13. Bowman, G.; Volez, V.; Pande, V.S. Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4–11.
14. Schütte, C.; Winkelmann, S.; Hartmann, C. Optimal control of molecular dynamics using Markov state models. *Math. Program. Ser. B* **2012**, *134*, 259–282.
15. Hartmann, C.; Schütte, C. Efficient rare event simulation by optimal nonequilibrium forcing. *J. Stat. Mech. Theor. Exp.* **2012**, doi:10.1088/1742-5468/2012/11/P11004.
16. Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M.E.; Noel, J.P.; Gruebele, M.; Kelly, J.W. Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10648–10653.
17. Kobitski, A.Y.; Nierth, A.; Helm, M.; Jäschke, A.; Nienhaus, G.U. Mg<sup>2+</sup> dependent folding of a Diels-Alderase ribozyme probed by single-molecule FRET analysis. *Nucleic Acids Res.* **2007**, *35*, 2047–2059.
18. Fischer, S.; Windshuegel, B.; Horak, D.; Holmes, K.C.; Smith, J.C. Structural mechanism of the recovery stroke in the Myosin molecular motor. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6873–6878.

19. Noé, F.; Krachtus, D.; Smith, J.C.; Fischer, S. Transition networks for the comprehensive characterization of complex conformational change in proteins. *J. Chem. Theory Comput.* **2006**, *2*, 840–857.
20. Ostermann, A.; Waschipky, R.; Parak, F.G.; Nienhaus, U.G. Ligand binding and conformational motions in myoglobin. *Nature* **2000**, *404*, 205–208.
21. Huisinga, W. Metastability of Markovian Systems a Transfer Operator Based Approach in Application to Molecular Dynamics. Ph.D Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, Berlin, Germany, 2001.
22. Bovier, A.; Eckhoff, M.; Gayraud, V.; Klein, M. Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc.* **2004**, *6*, 399–424.
23. Voronoi, M.G. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J. Reine Angew. Math.* **1908**, *134*, 198–287.
24. Sarich, M.; Noé, F.; Schütte, C. On the approximation quality of markov state models. *Multiscale Model. Simul.* **2010**, *8*, 1154–1177.
25. Sarich, M. Projected Transfer Operators. Ph.D. Thesis, Freie Universität Berlin, Berlin, Germany, 2011.
26. Sarich, M.; Schütte, C. Approximating selected non-dominant timescales by Markov state models. *Commun. Math. Sci.* **2012**, *10*, 1001–1013.
27. Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.
28. Faradjian, A.K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120*, 10880–10889.
29. Roebnitz, S. Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics. Ph.D. Thesis, Freie Universität Berlin, Berlin, Germany, 2008.
30. Djurdjevac, N.; Sarich, M.; Schütte, C. On Markov State Models for Metastable Processes. In Proceeding of the ICM 2010 as Invited Lecture, Hyderabad, India, 19–27 August 2010. Available online: <http://www.biocomputing-berlin.de/biocomputing/en/?cmd=publication> (accessed on 29 November 2010).
31. Horenko, I.; Dittmer, E.; Lankas, F.; Maddocks, J.; Metzner, P.; Schütte, C. Macroscopic dynamics of complex metastable systems: Theory, algorithms, and application to B-DNA. *J. Appl. Dyn. Syst.* **2008**, *7*, 532–560.
32. Weber, M.; Bujotzek, A.; Haag, R. Quantifying the rebinding effect in multivalent chemical ligand-receptor systems. *J. Chem. Phys.* **2012**, *137*, 054111.
33. Bowman, G.R.; Huang, X.; Pande, V.S. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **2009**, *49*, 197–201.
34. Schütte, C.; Sarich, M. *Metastability and Markov State Models in Molecular Dynamics. Modeling, Analysis, Algorithmic Approaches (Courant Lecture Notes No. 24)*; AMS: Providence, RI, USA, 2013.
35. Weber, M.; Fackeldey, K. Computing the minimal rebinding effect included in a given kinetics. *Multiscale Model. Simul.* **2013**. Available online: <http://www.zib.de/en/numerik/publications.html> (accessed on 1 December 2013).

36. Fleming, W.; Soner, H. *Controlled Markov Processes and Viscosity Solutions*; 2nd ed.; Springer: New York, NY, USA, 2005.
37. Oksendal, B. *Stochastic Differential Equations*; Springer: Heidelberg, Germany, 2003.
38. Pra, P.; Meneghini, L.; Runggaldier, W. Connections between stochastic control and dynamic games. *Math. Control Signals Syst.* **1996**, *9*, 303–326.
39. Kushner, H.; Dupuis, P. *Numerical Methods for Stochastic Control Problems in Continuous Time*; Springer Verlag: New York, NY, USA, 1992.
40. Braess, D. *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*; Cambridge University Press: Cambridge, UK, **2007**.
41. Banisch, R.; Hartmann, C. A meshfree discretization for optimal control problems. *SIAM J. Control Optim.* **2013**, submitted.
42. Sheu, S. Stochastic control and exit probabilities of jump processes. *SIAM J. Control Optim.* **1985**, *23*, 306–328.
43. Latorre, J.; Metzner, P.; Hartmann, C.; Schütte, C. A Structure-preserving numerical discretization of reversible diffusions. *Commun. Math. Sci.* **2011**, *9*, 1051–1072.
44. Chodera, J.D.; Elms, P.J.; Swope, W.C.; Prinz, J.H.; Marqusee, S.; Bustamante, C.; Noé, F.; Pande, V.S. A robust approach to estimating rates from time-correlation Functions. **2011**, arXiv:1108.2304.
45. Vanden-Eijnden, E. Transition path theory. In *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*; Ferrario, M., Ciccotti, G., Binder, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 703, pp. 353–493.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).