Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

CHRISTOF SCHÜTTE, ADAM NIELSEN AND MARCUS WEBER

Markov State Models and Molecular Alchemy

ZIB-Report 14-05 (March 2014, revised October 2014)

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 and Molecular Alchemy

Christof Schütte[a,b**], Adam Nielsen[b**†], Marcus Weber[b***‡]
[a] *Institute for Mathematics, Freie Universität Berlin, Germany*
[b] *Zuse Institute Berlin (ZIB), Germany*

Abstract

In recent years, Markov State Models (MSMs) have attracted a considerable amount of attention with regard to modelling conformation changes and associated function of biomolecular systems. They have been used successfully, e.g., for peptides including time-resolved spectroscopic experiments, protein function and protein folding, DNA and RNA, and ligand-receptor interaction in drug design and more complicated multivalent scenarios. In this article, a novel reweighting scheme is introduced that allows to construct an MSM for certain molecular system out of an MSM for a similar system. This permits studying how molecular properties on long timescales differ between similar molecular systems without performing full molecular dynamics simulations for each system under consideration. The performance of the reweighting scheme is illustrated for simple test cases, including one where the main wells of the respective energy landscapes are located differently and an alchemical transformation of butane to pentane where the dimension of the state space is changed.

1 Introduction

Applications in modern biochemistry, molecular medicine, or pharmacy demand for numerical simulations of large biomolecular systems in atomic representation aimed at a detailed understanding of relations between molecular structures, dynamical behavior and function. The processes constituting molecular function typically happen on timescales many orders of magnitude, say 10-15 orders, longer than the typical time steps of the simulation. Despite all progress in high performance, computing direct numerical simulation on such timescales often still is infeasible. Hence, there is an increasing need for coarse graining schemes to accurately capture the long-term kinetics of a molecular system. The last decade has seen a manifold of approaches to coarse graining molecular dynamics. One of these approach is *Markov State Modelling* in which the kinetics of a

** Email: Christof.Schuette@fu-berlin.de
†** Email: nielsen@zib.de
‡*** Email: weber@zib.de

molecular system is described by a Markov jump process or Markov chain with the dominant metastable conformations of a molecular system as Markov states [17, 18, 15]. In recent years Markov State Modelling has been applied with striking success to many different molecular systems like peptides including time-resolved spectroscopic experiments [3, 14, 9], proteins and protein folding [5, 11, 2], DNA [8], and ligand-receptor interaction in drug design [7, 4] and more complicated multivalent scenarios [23, 19].

Despite the tremendous progress in Markov State Models (MSMs) in recent years, one still has to put considerable effort into the construction of a MSM: a large number of short or medium-long molecular dynamics trajectories has to be generated for the specific molecular system and its precise environmental parameters (like temperature, solvent details, or pH value) of interest. However, in many applications one is not only interested in a specific system or in specific environmental parameters. For example, when designing functional molecules like in ligand design, one wants to compare the kinetic behavior and related functions of a variety of molecules (screening), and/or one needs to understand the influence of environmental parameters on the binding affinity. In order to answer these questions with techniques presently available, one would have to construct individual MSMs for each molecular species / parameter value of interest which is prohibitive in most practical cases. In such cases, *molecular alchemy* [10] would be helpful: In silico molecular alchemy starts from a molecular system M_A of which a certain property $P(M_A)$ is known or has been computed preliminarily and computes the same property $P(M_B)$ for a different molecular system, M_B, by successively transforming M_A into M_B during the simulation or by reweighting the simulation for M_A due to the transformation $M_A \to M_B$.

In this article, a computational scheme for in silico molecular alchemy in combination with MSMs is presented. We will assume that an MSM has been constructed from MD trajectory information for the specific molecular system M_A and precise environmental parameters, p. By using exponential reweighting techniques in path space, we will show how to compute the MSM for another molecular system and/or changed environmental parameters (M_B, p') solely based on the MD trajectories of the (M_A, p)-simulations. In this MSM reweighting scheme, each of the (M_A, p)-trajectories gets a new weight which enters into the computation of the transition matrix of the reweighted MSM. In theory, the reweighting procedure is *exact* so that the (M_B, p')-MSM can be constructed without a single (M_B, p')-simulation. In fact, we will demonstrate that this approach allows to handle even critical cases like differing numbers of dominant conformations, or different dimensions of state space. However, if the molecular system (M_A, p) is "too different" from (M_B, p') then the reweighting will require very many original MD trajectories to converge due to inefficient sampling caused by large variance as typical for all reweighting schemes if initial and target structure are too far apart.

This reweighting technique differs essentially from the more traditional sampling approaches used in molecular simulation. Instead of sampling the stationary distribution by clever trajectories and reweighting afterwards, we try to

avoid the sampling of new trajectories and obtain a new Markov State Model only by reweighting pre-assembled trajectories, even the case of other dimensions is possible (which is not the case for sophisticated methods like, e.g., nested sampling [20]). Therefore, our approach is different from the umbrella sampling [22] which uses penalty potentials in order to control the behavior of a trajectory, and is different from the Wang-Landau sampling which adjusts the potential stepwise by realizing new trajectories. Also our approach differs from the histogram reweighting [1] and replica exchange [16] (and its recent optimization [13, 6]), because, despite the fact that we only use pre-assembled trajectories, we do not reweight the stationary distributions with space weights, but rather weights for each single trajectory. Even further, this approach does not only achieve an approximation of the stationary distribution, but an estimation of the complete Markov State Model.

The article is structured as follows. First, the background of Markov State Modelling is introduced in Sec. 2. Next, the theory behind the reweighting scheme is discussed in Sec. 3. Finally, in Sec. 4, the algorithmic realization of the reweighting scheme is sketched and illustrative numerical experiments demonstrate the performance of the reweighting scheme for splitting conformations and state space different dimensions.

2 Markov State Models (MSMs)

We consider diffusive molecular dynamics,

$$dx_t = -\nabla_q V(x_t)dt + \sigma dB_t, \qquad \sigma^2 = 2\beta^{-1}Id, \tag{1}$$

where $F_{\text{int}} = \sigma\dot{B}_t$ denotes the (interal) forcing given by a $3N$-dimensional Brownian motion W_t, and V the engery landscape associated with the molecular system under consideration.

The process x_t admits a unique, positive *invariant probability measure* μ to which it is *ergodic*. This invariant density is absolutely continuous wrt. the Lebesgue measure and given by the density $\propto \exp(-\beta V(x))$ that we for convenience also denote by

$$\mu(x) = \frac{1}{Z}\exp(-\beta V(x)), \qquad Z = \int \exp(-\beta V(x))dx,$$

For an arbitrary complete decomposition $\{A_1, \ldots, A_m\}$ of state space into m disjoint sets we define the *transition matrix* to be the $m \times m$ matrix T with entries

$$T_{ij} = \mathbf{P}_\mu[x_t \in A_j \mid x_0 \in A_i], \tag{2}$$

where $\mathbf{P}_\mu$ indicates that X_0 is distributed due to μ. It is a stochastic matrix that describes the transition probabilities between the sets of the decomposition on time scale t in *equilibrium*. Its entries can alternatively be computed by

$$T_{ij} = \frac{1}{\mu(A_i)}\,\mathbf{E}_\mu\Big(\mathbf{1}_{A_i}(x_0)\mathbf{1}_{A_j}(x_t)\Big),$$

where $\mathbf{1}_A$ denotes the indicator function of the set A [18]. In the following, we assume that we already have identified appropriate sets $\{A_1, \ldots, A_m\}$ such that the associated transition matrix T forms a meaningful MSM *and* that we have already computed the entries of T up to sufficient accuracy by using sampling paths (i.e. trajectories) of the process $\{x_t\}$.

3 MSM reweighting

Now we are interested to understand the effect of a change of force in the SDE (1) on the entries of the MSM transition matrix T. This change of force can result from changing the environmental parameters of the simulation or from replacing groups of atoms of the molecular system by other ones. Even the case of introducing additional atoms is included as will be illustrated in Sec. 4.3.

Girsanov transformation

To this end, let $x_t = x_t(\omega)$ and $X_t = X_t(\omega)$ be the solutions on some probability space (Ω, Σ, P) of the stochastic differential equations

$$dx_t = -\nabla V(x_t)dt + \sigma dB_t \tag{3a}$$
$$dX_t = -(\nabla V(X_t) + \nabla U(X_t))dt + \sigma dB_t \tag{3b}$$

and deterministic initial conditions

$$x_0(\omega) = X_0(\omega) = x \quad \text{(almost surely)}\,.$$

Define $\xi_t \in \mathbf{R}^n$ by

$$\xi_t = \sigma^{-1}\nabla U(x_t) = \sqrt{\frac{\beta}{2}} \cdot \nabla U(x_t).$$

It follows from the Girsanov theorem [12, Thm. 8.6.8], sometimes also called *Cameron-Martin-Girsanov theorem* [21] that for

$$dQ := M_t dP$$

with

$$M_t := \exp\left(-\int_0^t \xi_s \cdot dB_s - \frac{1}{2}\int_0^t |\xi_s|^2 ds\right), \tag{4}$$

we get for any measurable set A

$$P[X_t \in A] = Q[x_t \in A],$$

which is identical to writing

$$\int \mathbf{1}_A(X_t(\omega))dP = \int \mathbf{1}_A(x_t(\omega))dQ.$$

In particular, we obtain

$$\begin{aligned}\mathbf{E}[\mathbf{1}_A(X_t)] &= \int \mathbf{1}_A(X_t(\omega))dP \\ &= \int \mathbf{1}_A(x_t(\omega))dQ \\ &= \int \mathbf{1}_A(x_t(\omega))M_t(\omega)dP \\ &= \mathbf{E}[M_t\mathbf{1}_A(x_t)]\end{aligned} \tag{5}$$

for any measurable set A.

Updating MSM transition probabilities

Now let T^Q denote the transition matrix associated with $\{X_t\}$ for the same sets $\{A_1, \ldots, A_m\}$, and μ_Q the associated invariant measure. Then, our result yields that

$$\begin{aligned}T_{ij}^Q &= \frac{1}{\mu_Q(A_i)}\,\mathbf{E}_{\mu_Q}\Big(\mathbf{1}_{A_i}(X_0)\mathbf{1}_{A_j}(X_t)\Big) \qquad (6)\\ &= \frac{1}{\mu_Q(A_i)}\int \mathbf{1}_{A_i}(x)\cdot\mathbf{E}_x\left[\mathbf{1}_{A_j}(x_t)\exp\left(-\int_0^t \xi_s\cdot dB_s - \frac{1}{2}\int_0^t |\xi_s|^2 ds\right)\right]\mu_Q(x)dx.\end{aligned}$$

We have

$$\begin{aligned}\mu_Q(x) &= \frac{1}{Z_Q}\exp\Big(-\beta(V(x)+U(x))\Big) = \frac{Z}{Z_Q}\mu(x)\exp(-\beta U(x)),\\ Z_Q &= \int \exp\Big(-\beta(V(x)+U(x))\Big)dx = Z\cdot\mathbf{E}_\mu(e^{-\beta U}).\end{aligned}$$

such that

$$\begin{aligned}T_{ij}^Q &= \frac{1}{\mu_Q(A_i)}\int_{A_i} w_j(t,x)\,g(x)\,\mu(x)dx, \qquad (7)\\ w_j(t,x) &= \mathbf{E}_x\left[\mathbf{1}_{A_j}(x_t)\exp\left(-\int_0^t \xi_s\cdot dB_s - \frac{1}{2}\int_0^t |\xi_s|^2 ds\right)\right] \qquad (8)\\ \xi_s &= \sqrt{\frac{\beta}{2}}\cdot\nabla U(x_s)\\ g(x) &= \frac{e^{-\beta U(x)}}{\mathbf{E}_\mu(e^{-\beta U})},\end{aligned}$$

with $\mu_Q(A_i) = \int_{A_i} g(x)\,\mu(x)dx$. Consequently, based on the trajectory information that was gained to compute T_{ij}, we in principle can also compute T_{ij}^Q.

Note that for

$$C_{ij} = \int_{A_i} w_j(t,x)\,\tilde{g}(x)\,\mu(x)dx$$

with

$$\tilde{g}(x) = e^{-\beta U(x)}$$

we obtain

$$\sum_{j=1}^{m} C_{ij} = \int_{A_i} \tilde{g}(x)\,\mu(x)dx$$

and, therefore,

$$\frac{C_{ij}}{\sum_{j=1}^m C_{ij}} = \frac{C_{ij}}{\sum_{j=1}^m C_{ij}} \frac{\frac{1}{c}}{\frac{1}{c}} = T^Q_{ij}$$

for $c = \mathbf{E}_\mu(e^{-\beta U})$

Linearization

Denote by T^P the transition matrix associated with $\{x_t\}$ for the sets $\{A_1, \ldots, A_m\}$. Let us now consider the case of a small change in potential, i.e., $U = \varepsilon W$ with a small $\varepsilon > 0$. Then Taylor expansion around $\varepsilon = 0$ for the weighting factor $w(t,x)$ yields the update formula

$$\begin{aligned} T^Q_{ij} &= T^P_{ij} + \varepsilon\, L^P_{ij} + \mathcal{O}(\varepsilon^2), \\ L^P_{ij} &= \frac{1}{\mu_Q(A_i)} \sqrt{\frac{\beta}{2}} \int_{A_i} \mathbf{E}_x \left[\mathbf{1}_{A_j}(x_t) \int_0^t \nabla W(x_s)\, dB_s \right] g(x)\,\mu(x)\,dx, \\ \mu_Q(A_i) &= \int_{A_i} g(x)\,\mu(x)dx. \end{aligned} \tag{9}$$

Here L^Q_{ij} still depends on ε via g and μ_Q; however, we do not linearize these factors for two reasons: (1) linearization may destroy the normalization of the measure and (2) these factors can be computed pointwise and thus rather efficiently. By exchanging groups of atoms in the molecular system under consideration one gets a specific update potential U. Then $U = W$ and evaluation of C^P gives the sensitivity of the MSM matrix to the exchange.

Generalization

Next we consider the Langevin equation with position q and associated momenta p:

$$\begin{aligned} dq &= M^{-1}p \\ dp &= \Big[-\nabla V(q) + \gamma p \Big]\, dt + \sigma dB_t, \end{aligned}$$

where γ is the friction coefficient. The state of the system is $x = (q,p)$ with invariant measure

$$\mu(x) = \frac{1}{Z} \exp\left(-\frac{\beta}{2} \Big[p^T M^{-1} p + V(q) \Big] \right), \qquad \beta = \frac{2\gamma}{\sigma^2}.$$

In this case the above reweighting argument for replacing V by $V+U$ goes as before. This time the definition of ξ is

$$\begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix} \begin{pmatrix} \xi_q \\ \xi_p \end{pmatrix} = \begin{pmatrix} 0 \\ -\nabla U(q) \end{pmatrix}.$$

Thus our reweighting formula (7) remains unchanged, now with

$$\begin{aligned} w_j(t,x) &= \mathbf{E}_x \left[\mathbf{1}_{A_j}(x_t) \exp\left(-\int_0^t \xi_{p,s} \cdot dB_s - \frac{1}{2}\int_0^t |\xi_{p,s}|^2 ds \right) \right] \\ \xi_{p,s} &= \sqrt{\frac{\beta}{2}} \cdot \nabla U(q_s). \end{aligned}$$

Further generalization to other forms of molecular dynamics, like thermostatted MD, is possible by incorporating the respective environmental forces as an stochastic forcing.

4 Algorithmic Realization and Numerical Experiments

In the following, we compare different approximations of the transition matrix T^Q related to (3b). One form of approximation of T^Q is through direct computation, i.e., based on (6) using trajectories of (3b); we will denote this approximation by $T^{Q,dir}$. The other one results from the reweighting scheme based on trajectories of (3a), denoted by $T^{Q,reweighted}$. We explain both approximation types now in detail. In the following, X_t, x_t denote d-dimensional random variables.

Direct computation To gain $T^{Q,dir}$, we compute a long trajectory $(X_i)_{i=0,\dots,n-1}$ for the perturbed dynamics (3b) by performing n timesteps of size dt using the Euler-Maruyama discretization

$$X_{i+1} = X_i - (\nabla V(X_i) + \nabla U(X_i))\, dt + \sigma\sqrt{dt}\,\eta_i$$

of (3b), where $\eta_i = (\eta_i^1, \dots, \eta_i^d)$ are independent d-dimensional random variables distributed due to the standard normal distribution. This trajectory is chopped into pieces of length l yielding M subtrajectories $(X_i^k)_{i=1,\dots,l} := (X_{lk}, \dots, X_{l(k+1)-1})$ for $k = 0, \dots, M-1$. If the trajectory is long enough, it can be assumed that the points $X_1^0, \dots, X_1^{M-1}$ are distributed according to μ_Q and we can calculate $T^{Q,dir}$ by

$$C_{ij}^D = \sum_{k=0}^{M-1} \mathbf{1}_{A_i}(X_1^k)\, \mathbf{1}_{A_j}(X_l^k)$$

and

$$T_{ij}^{Q,dir} = \frac{C_{ij}^D}{\sum_{i=1}^m C_{ij}^D}. \tag{10}$$

Reweighted computation To gain $T^{Q,reweighted}$, we compute a long trajectory $(x_i)_{i=0,\dots,n-1}$ for the unperturbed dynamics (3a) by performing n timesteps of size dt using the Euler-Maruyama discretization

$$x_{i+1} = x_i - (\nabla V(x_i))\, dt + \sigma\sqrt{dt}\, \eta_i$$

of (3a), where $\eta_i = (\eta_i^1, \dots, \eta_i^d)$ are independent d-dimensional random variables distributed due to the standard normal distribution. Again, this trajectory is chopped into pieces of length l yielding M subtrajectories $(x_i^k)_{i=1,\dots,l} := (x_{lk}, \dots, x_{l(k+1)-1})$ for $k = 0, \dots, M-1$. If the trajectory is long enough, it can be assumed that the points $x_1^0, \dots, x_1^{M-1}$ are distributed according to μ. Now, we have to approximate for each subtrajectory $(x_i^k)_{i=1,\dots,l}$ the term $M_t(x_l^k)$ with $t = l \cdot dt$. This is done as follows. First, for

$$\mathcal{R} = \int_0^t \xi_s\, dB_s + \frac{1}{2}\int_0^t |\xi_s|^2 ds = \sum_{i=1}^d \left(\int_0^t \xi_s(i)\, dB_s^i \right) + \frac{1}{2}\int_0^t |\xi_s|^2 ds \tag{11}$$

we have $M_t = \exp(-\mathcal{R})$, where $B_s = \left(B_s^1, \dots, B_s^d\right)$ denotes the d-dimensional Brownian Motion with independent components. Each term $\int_0^t \xi_s(i)\, dB_s^i$ can be approximated by using the Euler-Maruyama discretization by

$$\int_0^t \xi_s(i)\, dB_s^i \approx r_l^i - r_0^i,$$

where

$$\begin{aligned} r_0^i &= x_1^k \\ r_{j+1}^i &= r_j^i + [\xi(r_j^i)(i)]\, \eta_{(kl+j)}^i \sqrt{dt} \end{aligned}$$

and

$$\xi(r) = \sigma^{-1} \nabla U(r).$$

Therefore, for each trajectory $(x_i^k)_{i=1,\dots,l}$ we calculate the weight w_k by

$$\begin{aligned} r_k &= \sum_{i=1}^d (r_l^i - r_0^i) + \frac{1}{2}\sum_{i=1}^l |\xi(x_i^k)|^2\, dt \\ w_k &= \exp(-r_k). \end{aligned} \tag{12}$$

Finally, we can compute $T^{Q,reweighted}$ by

$$C_{ij}^w = \sum_{k=0}^{M-1} \mathbf{1}_{A_i}(x_1^k)\, \mathbf{1}_{A_j}(x_l^k)\, w_k\, \tilde{g}(x_1^k) \tag{13}$$

and

$$T_{ij}^{Q,reweighted} = \frac{C_{ij}^w}{\sum_{i=1}^m C_{ij}^w}.$$

It should be obvious that we also can use the trajectory $(x_i)_{i=0,\ldots,n-1}$ for construction of an approximation $T^{P,dir}$ of T^P in direct analogy of the construction of $T^{Q,dir}$ based on $(X_i)_{i=0,\ldots,n-1}$.

Three remarks may be in order:

- The trajectory $(X_i)_{i=0,\ldots,n-1}$ for the perturbed energy landscape $V+U$ is computed solely for allowing comparisons; in real-world applications of the reweighting scheme only trajectory information for the unperturbed energy landscape V will be computed and the perturbing force field ∇U is evaluated solely as part of the weight calculations.
- The decomposition of the available long trajectory into subtrajectories makes it obvious how to use the reweighting scheme if only an ensemble of short trajectories instead of one long trajectory is given.
- It is essential that the random vector η_i which is used to compute (x_i) is also used to compute the weights M_t.

4.1 Double Well Potential

We consider the two one-dimensional potentials

$$V(x) = (x^2-1)^2, \quad U(x) = ax,$$

and want to see whether our reweighting formula allows to reproduce the tilting of the double well potential V by the linear perturbation U. For our experiments we chose $a = -0.75$ which results in the potentials shown in Fig. 1.

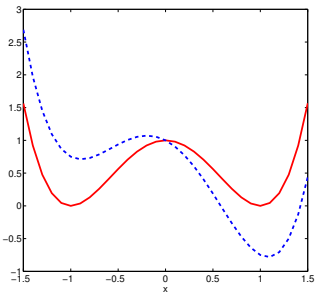

Figure 1: Double well potentials V (red, solid) and $V+U$ (blue, dashed).

Setting $\beta = 2.5$, we apply the above scheme for the constrcution of $T^{Q,reweighted}$ and compute a trajectory with $n = 4 \cdot 10^7$ timesteps of size $dt = 0.001$ which we chop into pieces of length $l = 400$, yielding $M = 100.000$ subtrajectories $(x_i^k)_{i=1\ldots l}$ $k = 0, \ldots, M-1$. This long-term trajectory samples the interval $S = [-1.6, 1.6]$. Next, we compute the MSM matrices $T^{Q,reweighted}$ and $T^{P,dir}$ for $t = 0.4$ and based on the following complete decomposition $A_1, \ldots, A_m$ of S:

eigenvalues	$T^{P,dir}$	$T^{Q,dir}$	$T^{Q,reweighted}$	$T^{Q,linear}$	$T^{Q,no-weights}$
λ_3	0.227	0.189	0.190	0.200	0.221
λ_2	0.947	0.886	0.884	0.895	0.949

Table 1: Results for the double well potential: Second and third eigenvalues of the different MSM matrix approximation introduced in the text.

$A_i = [x_i, x_i + \Delta x)$, $x_i = -1.6 + (i-1)\Delta x$ for $i = 1, \ldots, m = 32$ and $\Delta x = 0.1$. For further comparison, we compute the following alternative approximations of T^Q:

- We ignore the reweighting factor w_k in (13), i.e., we use (13) with $w_k = 1$ to get $T^{Q,no-weights}$.
- We replace the reweighting factor w_k in (13) by their linearized counterparts as of (9) and get $T^{Q,linear} = T^P_{ij} + \varepsilon L^Q_{ij}$.

For the sole sake of comparison, we compute $T^{Q,dir}$ based on a trajectory of the perturbed dynamics (3b) with same length and sampling parameters, and for the same complete decomposition.

Table 1 gives the second and third eigenvalues of the respective transition matrices. We observe

- that our the leading eigenvalues of the reweighted transition matrix $T^{Q,reweighted}_{ij}$ are almost identical with the dominant ones of $T^{Q,dir}$; this demonstrates the validity of our formula (7),
- that the dominant eigenvalues of the linearization $T^{Q,linear}$ are close (but not very close) to those of the full reweighted transition matrix $T^{Q,reweighted}_{ij}$,
- that $T^{Q,no-weights}$ has dominant eigenvalues similar to those of T^P but rather different from those of $T^{Q,reweighted}_{ij}$ and $T^{Q,dir}$; this shows that the weighting factors w_k have a decisive impact on the transition probabilities,

Figure 2 shows the invariant measure of the respective approximations, compared to the measures computed based on the exact formula $\mu_P = \exp(-\beta V)/Z$ and $\mu_Q = \exp(-\beta(V+U))/Z_Q$. We observe that the invariant measure of the respective approximations are almost identical with small deviation resulting from the sampling and discretization errors.

4.2 Splitting a metastable set

Here we stick with the double well potential considered above but choose U differently,

$$U(x) = a \exp\Big(-\frac{1}{2s^2}(x+1)^2 \Big),$$

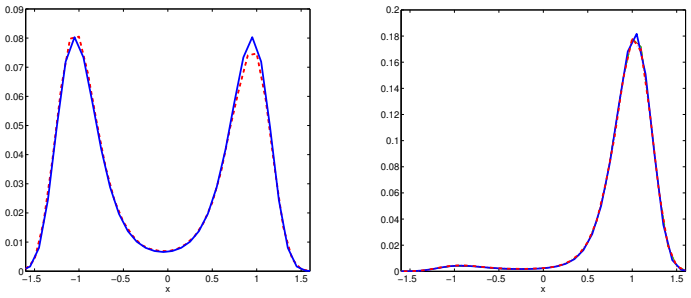

Figure 2: Results for the double well potential. Left: Invariant measures μ_P (blue, solid) compared to μ computed via $\mu = \mu T^P$ based on the approximate MSM matrix T^P (red, dashed). Right: Invariant measures μ_Q (blue, solid) compared to $\tilde{\mu}$ computed via $\tilde{\mu} = \tilde{\mu} T^{Q,reweighted}$ based on the approximate MSM matrix $T^{Q,reweighted}$ (red, dashed).

with $a = 1.25$, and $s = 0.05$. This results in the situation shown in Fig. 3, where the left well of V is split by the peak of U such that $V + U$ has three instead of two metastable sets.

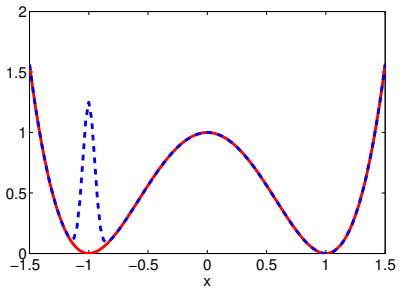

Figure 3: Split double well potential: Potentials V (red, solid) and $V + U$ (blue, dashed).

We perform the same algorithmic procedure as above (this time with $t = 0.15$ and $\Delta x = 0.05$), resulting in the MSM matrices $T^{P,dir}$, $T^{Q,dir}$, and $T^{Q,reweighted}$. The second and third eigenvalues of these matrices are shown in Table 2. We observe that the creation of the third metastable well induces the third eigenvalue to move closer to the second one. This is reproduced by the reweighting with sufficient (but not perfect) accuracy. The reason for the deviation between the third eigenvalues of $T^{Q,dir}$ and $T^{Q,reweighted}$ lies in the following sampling problem: Around $x = -1$ the perturbation potential U exhibits steep gradients. However, the second and third eigenvectors of $T^{Q,dir}$ and $T^{Q,reweighted}$ essentially agree, see Fig. 4.

eigenvalues	$T^{P,dir}$	$T^{Q,dir}$	$T^{Q,reweighted}$
λ_3	0.603	0.900	0.881
λ_2	0.993	0.991	0.990

Table 2: Results for the split double well potential: Second and third eigenvalues of the different MSM matrix approximation.

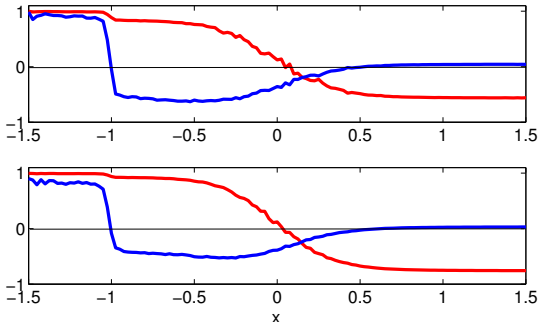

Figure 4: Results for the split double well potential: Second (red) and third (red) eigenvector of transition matrix $T^{Q,dir}$ (top panel) and $T^{Q,reweighted}$ (bottom panel).

4.3 From Butane to Pentane

In the following, it is shown how to apply the result for different dimensions. Consider the one dimensional, 2π-periodic, artificial potential $V_B \colon \mathbb{R} \to \mathbb{R}$ of the dihedral angles for butane given by

$$V_B(x) = a + b\,\cos(x) + c\,\cos^2(x) + d\,\cos^3(x)$$

with a=2.0567, b=-4.0567, c=0.3133, d=6.4267 and the two dimensional, 2π-periodic, artificial potential $V_P \colon \mathbb{R}^2 \to \mathbb{R}^2$ of the dihedral angles for pentane given by

$$V_P(x, y) = V_B(x) + V_B(y)$$

as shown in Figure 5.

To obtain the transition matrix from pentane by simulations of butane we use the Girsanov transformation with $V(x, y) = V_B(x)$ and $U(x, y) = V_P(x, y) - V_B(x)$. We still denote with T^Q the transition matrix associated with $\{X_t\}$ from (3b), which depends on V and U. If we choose $\sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_1 \end{pmatrix}$ with $\sigma_1^2 = 2\beta^{-1}$

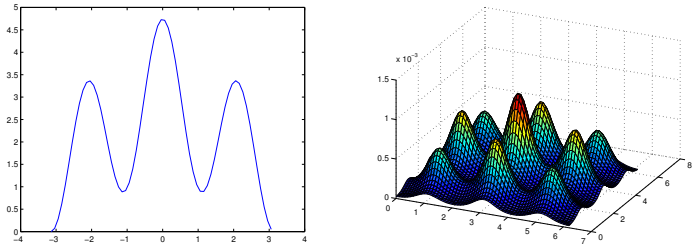

Figure 5: Left: Boltzmann distribution $\exp(-\beta V_B(x))$ on $[0, 2\pi)^2$. Right: Boltzmann distribution $\exp(-\beta V_P(x))$ on $[0, 2\pi)^2$.

and $\beta = 0.5$, then, when replacing $x_t = \begin{pmatrix} y_t \\ z_t \end{pmatrix}$, (3a) becomes

$$dy_t = \frac{\partial V_B(y_t)}{\partial x} + \sigma_1 dB_t^1 \tag{14}$$

$$dz_t = \sigma_1 dB_t^2. \tag{15}$$

This shows that y_t and z_t can be solved independently. Therefore, in case a representative trajectory for butane has already been computed, we only need one trajectory of the Brownian motion to compute the approximation $T^{Q,reweighted}$ for pentane. This Brownian motion is completely independent of y_t.

This time, we compute such a representative trajectory by performing $n = 4 \cdot 10^8$ timesteps of size $dt = 0.001$ using the Euler-Maryama discretization

$$y_{i+1} = y_i - \nabla V(y_i)\, dt + \sigma_1 \sqrt{dt}\, \eta_i^1,$$
$$z_{i+1} = z_i + \sigma_1 \sqrt{dt}\, \eta_i^2.$$

This yields two statistically independent discrete trajectory y_i, z_i, $i = 0, \ldots, 4 \cdot 10^8 - 1$ plotted in Figure 6.

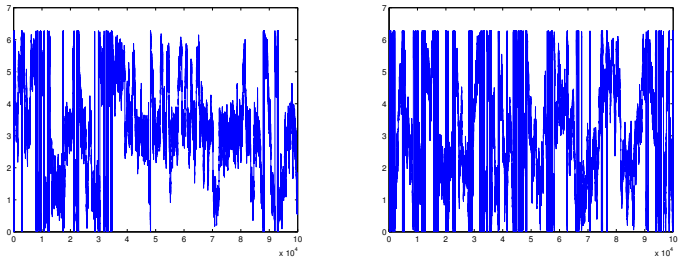

Figure 6: Left: Trajectory y_t mod 2π. Right: Trajectory z_t mod 2π.

This long trajectory is chopped into pieces of length $l = 400$, yielding $M = 10.000.000$ subtrajectories $(y_i^k)_{i=1,\ldots,l}$, $(z_i^k)_{i=1,\ldots,l}$, $k = 0, \ldots, M-1$.

eigenvalues	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
$T^{Q,weighted}$	0.952	0.947	0.946	0.941	0.902	0.896	0.895	0.890	0.648
$T^{Q,dir}$	0.952	0.952	0.946	0.946	0.906	0.901	0.900	0.895	0.643
$T^{P,dir}$	0.952	0.946	0.637	0.642	0.443	0.441	0.251	0.252	0.121

Table 3: From butane to pentane: First dominating eigenvalues of $T^{Q,reweighted}$, $T^{Q,dir}$ and $T^{P,dir}$, where $T^{P,dir}$ denotes the directly computed transition matrix for butane according to the process y_t from V_B by dividing $[0, 2\pi)$ into 30 sets.

We divide $[0, 2\pi)$ into 30 sets $A_i = [x_i, x_i + \Delta x)$, $i = 1, \ldots, 30$ with $x_i = (i-1)\Delta x$ and $\Delta x = 2\pi/30$. Then, we partition $[0, 2\pi)^2$ into 900 sets B_{ij} with $B_{ij} = A_i \times A_j$ for $i, j = 1, \ldots, 30$ and use the above scheme to construct $T^{Q,reweighted}$.

In addition we compute an analogous trajectory for pentane and construct $T^{Q,dir}$ based on the same complete partition (B_{ij}), again just for the sake of comparison.

The eigenvalues given in Table 3 and the eigenvector for eigenvalue $\lambda = 1$ given in Figure 7 show that the weighted transition matrix $T^{Q,reweighted}$ is a good approximation of $T^{Q,dir}$.

From Butane to Pentane with MD in atomic resolution

In the preceding section, we have seen that we can go from butane to pentane simply by simulating a Brownian motion on the interval $[0, 2\pi]$. For MD in full atomic resolution, a Brownian motion on $\mathbb{R}^{3n}$ will not be helpful. However, we can proceed as follows. First, we need to split the molecule in two parts that coincide in one single atom, see Figure 8. Then, for each part we can simulate an independent MD trajectory in full atomic resolution. If we merge both trajectories together, the resulting trajectory can be seen as the outcome of the original potential without taking into account the interactions between both parts. To be more precise, let us label the position in state space of the atoms from pentane at time step i by $x_i^1, \ldots, x_i^{17} \in \mathbb{R}^3$. Then we are going to fix the yellow atom 13 in state space, for example $x_i^{13} = (0, 0, 0)$ for all $i \in \mathbb{N}$ (see Figure 8). The first part of pentane which consists now of 12 atom positions can be evaluated by a force field V_1 which only takes into account the interactions of the first 13 atoms where the 13th atom is fixed:

$$\begin{pmatrix} x_{i+1}^1 \\ \vdots \\ x_{i+1}^{12} \end{pmatrix} = -\nabla V_1 \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^{12} \end{pmatrix} dt + \sigma \begin{pmatrix} dB_i^1 \\ \vdots \\ dB_i^{12} \end{pmatrix}.$$

Analogously we can evaluate the second part by

$$\begin{pmatrix} x_{i+1}^{14} \\ \vdots \\ x_{i+1}^{17} \end{pmatrix} = -\nabla V_2 \begin{pmatrix} x_i^{14} \\ \vdots \\ x_i^{17} \end{pmatrix} dt + \sigma \begin{pmatrix} dB_i^{14} \\ \vdots \\ dB_i^{17} \end{pmatrix}.$$

Altogether, denoting z as

$$z_i := \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^{12} \\ x_i^{14} \\ \vdots \\ x_i^{17} \end{pmatrix} \quad \text{and} \quad V(z_i) := V_1 \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^{12} \end{pmatrix} + V_2 \begin{pmatrix} x_i^{14} \\ \vdots \\ x_i^{17} \end{pmatrix}$$

we can consider z as the solution of the stochastic differential equation

$$dz_t = -\nabla V(z_t)\, dt + \sigma dB_t. \tag{16}$$

If $U(z)$ denotes the potential that takes into account the interactions between $(x^1, \ldots, x^{12})$ and $(x^{14}, \ldots, x^{17})$, we can reformulate the problem how to construct the MSM of

$$dz_t = -\nabla(V(z_t) + U(z_t))\, dt + \sigma dB_t$$

if only trajectories of (16) are available. This can be solved with our the machinery already presented above. Calculating the weights will be much more efficient than calculation new trajectories, because we only have to consider the interactions between both molecular parts for calculating the corresponding weights.

5 Conclusion

In this article we have assumed that an MSM for a certain molecular system has already been constructed based on a long-term molecular trajectory (or an ensemble of shorter ones). We have shown how this trajectory can be re-used to also compute the MSM of a perturbed molecular system by reweighting. The underlying reweighting scheme has been derived for diffusive molecular dynamics. including Langevin dynamics. As a by-product linearization of the reweighting provides hints for a sensitivity analysis of molecular systems according to small force field or parameter changes. We have illustrated the performance of the reweighting scheme for simple test cases including one where the perturbed energy landscape exhibits additional wells and an alchemical transformation of butane to pentane where the dimension of the state space is changed. In the numerical experiment from butane to pentane it turned out that for long trajectories one has to take a sufficiently small time step, in order to get acceptable

weights. The term r_k of the weight $w_k = \exp(-r_k)$ consists of two components: A stochastic integral and a classical positive integral along the trajectory, see Equation (11). The second term increases according to the length of the trajectory and the perturbation U, which causes the weights to tend to zero for long trajectories. However, the first term can reverse this effect, but may demand a small time step to obtain a sufficient approximation of the Ito integral.

The presented method shares the fundamental difficulties of all reweighting schemes: (1) If the perturbation of the molecular system is "too large" then the reweighting might yield inaccurate results since the reweighting renders most of the original trajectories statistically irrelevant, or (2) if the perturbation of the molecular force field is not local then the reweighting may be computationally very expensive since then the computation of the new trajectory weights along the trajectories require too expensive force field evaluations. Within these restrictions the application of the scheme in typical MD cases will produce considerably less computational effort than the from scratch construction of an additional MSM for the perturbed molecular system. This article outlines how to perform such realistic MD applications while being restricted to the fundamentals; details of an algorithmic realization presently are topic of further investigations.

Acknowledgement Financial support from the DFG research center MATHEON is gratefully acknowledged. We would also like to thank Wei Zhang and Carsten Hartmann for fruitful discussions about path measures.

References

[1] Ferrenberg A.M. and R.H. Swendsen. *Physical Review Letters*, 63:1195–1198, 1989.

[2] G.R. Bowman, K.A. Beauchamp, G. Boxer, and V.S. Pande. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, 131(12):124101, September 2009.

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

[4] A. Bujotzek and M. Weber. Efficient Simulation of Ligand-Receptor Binding Processes Using the Conformation Dynamics Approach. *J. Bioinf. Comp. Bio.*, 7(5):811–831, 2009.

[5] 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.

[6] J. D. Chodera, W. C. Swope, F. Noé, J.-H. Prinz, and V. S. Pande. Dynamical reweighting: Improved estimates of dynamical properties from simulations at multiple temperatures. *submitted to J. Phys. Chem.*, 2010.

[7] Peter Deuflhard and Christof Schütte. Molecular conformation dynamics and computational drug design. In J. M. Hill and R. Moore, editors, *Applied Mathematics Entering the 21st Century. Proc. ICIAM 2003, Sydney, Australia*, pages 91–119, 2004.

[8] 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.*, 7(2):532–560, 2009.

[9] Bettina Keller, Andreij Y. Kobitski, Uli G. Nienhaus, and Frank Noé. Analysis of single molecule fret trajectories with hidden markov models. *Biophys. J.*, 2011.

[10] Alison Mitchell. Molecular alchemy. *Nature Reviews Molecular Cell Biology*, 1, 164, 2000.

[11] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T.R. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106:19011–19016, 2009.

[12] B.K. Øksendal. *Stochastic differential equations: an introduction with applications*. Springer Verlag, Berlin, Heidelberg, New York, 6th ed. edition, 2003.

[13] J.-H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith, and F. Noé. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.*, 134:244108, 2011.

[14] J.-H. Prinz, B. Keller, and F. Noé. Probing molecular kinetics with Markov models: Metastable states, transition pathways and spectroscopic observables. *Phys Chem Chem Phys.*, 13(38):16912–27, 2011.

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

[16] Swendsen RH and Wang JS. Replica monte carlo simulation of spin glasses. *Physical Review Letters*, 57:2607–2609, 1986.

[17] 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. Special Issue on Computational Biophysics.

[18] Ch. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics. Modeling, Analysis, Algorithmic Approaches.* Courant Lecture Notes, No. 24. AMS, 2013.

[19] M. Shan, K.E. Carlson, A. Bujotzek, A. Wellner, R. Gust, M. Weber, J.A. Katzenellenbogen, and R. Haag. Nonsteroidal Bivalent Estrogen Ligands – An Application of the Bivalent Concept to the Estrogen Receptor. *ACS Chem. Biol.*, 8(4):707–715, 2013.

[20] J. Skilling. Bayesian analysis 1. pages 833–860, 2006.

[21] D.W. Stroock and S.R.S. Varadhan. *Multidimensional Diffusion processes.* Springer, Berlin, Heidelberg, 2006.

[22] G.M. Torrie and Valleau. *J.-P. Jounral of Computational Physics*, 23:187–199, 1977.

[23] M. Weber, A. Bujotzek, and R. Haag. Quantifying the rebindig effect in multivalent chemical ligand-receptor systems. *J. Chem. Phys.*, 137(5):054111, 2012.

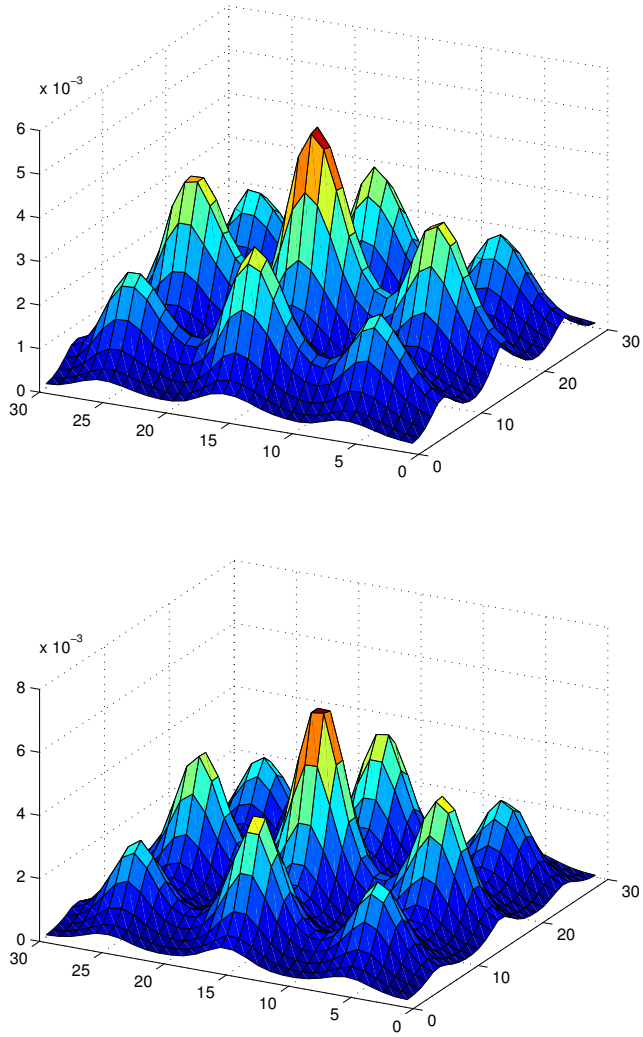

Figure 7: Top: Invariant measures μ_1 computed via $\mu_1 = \mu_1 T^{Q,reweighted}$. Bottom: Invariant measures μ_2 computed via $\mu_2 = \mu_2 T^{Q,dir}$.

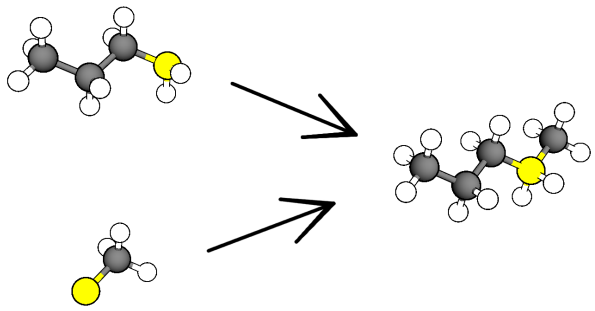

Figure 8: Left: Splitting pentane in two parts. Right: Merged parts.