Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias

Feliks Nüske, Hao Wu, Jan-Hendrik Prinz, Christoph Wehmeyer, Cecilia Clementi, and Frank Noé

Articles you may be interested in

Reconciling transition path time and rate measurements in reactions with large entropic barriers
The Journal of Chemical Physics **146**, 071101071101 (2017); 10.1063/1.4977177

Identification of simple reaction coordinates from complex dynamics
The Journal of Chemical Physics **146**, 044109044109 (2017); 10.1063/1.4974306

Perspective: Dissipative particle dynamics
The Journal of Chemical Physics **146**, 150901150901 (2017); 10.1063/1.4979514

Exploring high dimensional free energy landscapes: Temperature accelerated sliced sampling
The Journal of Chemical Physics **146**, 094108094108 (2017); 10.1063/1.4977704

Rate constants from instanton theory via a microcanonical approach
The Journal of Chemical Physics **146**, 074105074105 (2017); 10.1063/1.4976129

Beyond histograms: Efficiently estimating radial distribution functions via spectral Monte Carlo
The Journal of Chemical Physics **146**, 094107094107 (2017); 10.1063/1.4977516

Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias

Feliks Nüske,[1,a)] Hao Wu,[1] Jan-Hendrik Prinz,[1] Christoph Wehmeyer,[1] Cecilia Clementi,[2] and Frank Noé[1,a)]

[1]*Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany*
[2]*Center for Theoretical Biological Physics and Department of Chemistry, Rice University, Houston, Texas 77005, USA*

Many state-of-the-art methods for the thermodynamic and kinetic characterization of large and complex biomolecular systems by simulation rely on ensemble approaches, where data from large numbers of relatively short trajectories are integrated. In this context, Markov state models (MSMs) are extremely popular because they can be used to compute stationary quantities and long-time kinetics from ensembles of short simulations, provided that these short simulations are in "local equilibrium" within the MSM states. However, over the last 15 years since the inception of MSMs, it has been controversially discussed and not yet been answered how deviations from local equilibrium can be detected, whether these deviations induce a practical bias in MSM estimation, and how to correct for them. In this paper, we address these issues: We systematically analyze the estimation of MSMs from short non-equilibrium simulations, and we provide an expression for the error between unbiased transition probabilities and the expected estimate from many short simulations. We show that the unbiased MSM estimate can be obtained even from relatively short non-equilibrium simulations in the limit of long lag times and good discretization. Further, we exploit observable operator model (OOM) theory to derive an unbiased estimator for the MSM transition matrix that corrects for the effect of starting out of equilibrium, even when short lag times are used. Finally, we show how the OOM framework can be used to estimate the exact eigenvalues or relaxation time scales of the system without estimating an MSM transition matrix, which allows us to practically assess the discretization quality of the MSM. Applications to model systems and molecular dynamics simulation data of alanine dipeptide are included for illustration. The improved MSM estimator is implemented in PyEMMA of version 2.3. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4976518]

I. INTRODUCTION

Ensemble approaches, where many fairly short simulations are produced in parallel or on distributed computer architectures, are widely used in order to characterize the thermodynamics and kinetics of large biological macromolecules. Markov state models (MSMs)[1–3] have become standard tools for the analysis of such data sets generated by molecular dynamics (MD) simulations.[4–13] An MSM provides a simplified model of the underlying Markov process, which is continuous in both time and space, by a discrete time Markov chain on finitely many states. These states are defined by partitioning the continuous state space into finitely many disjoint sets. Time is discretized by choosing a discrete time step, called the *lag time*, and the full process is replaced by a snapshot process that only keeps track of the discrete state visited at the discrete time steps, discarding any time information in between and any spatial information within the discrete sets. The quality of this approximation critically depends on the choice of both discretization and lag time.[14] One of the

strengths of Markov models is that the simulations used to construct them do not necessarily need to sample from the global equilibrium distribution, as only conditional transition probabilities between the states are required.[4] In particular, at least in principle, these transition probabilities can be obtained without bias from simulations started out of local equilibrium in each state which only run for the length of a single lag time step. However, it is much more practical to produce simulations that are longer than one lag time and estimate MSMs by counting transitions along these trajectories. Even if the simulations are started out of local equilibrium, the distribution deviates from local equilibrium over time until global equilibrium is restored. The estimation of transition probabilities is therefore subjected to a bias.[1] In order to keep the bias small, it must be assumed that local equilibrium is approximately restored after every time step.

The effect of the initial distribution onto the MSM quality or even the justification of using an MSM for data analysis has been controversially discussed, and this issue has not been resolved yet. At least three ideas have been discussed as follows:[15] (1) This effect exists,[1] but may be small and can be ignored in practice. (2) We can reduce the effect of non-equilibrium starting points by discarding the first bit of

---

a)Electronic addresses: feliks.nueske@fu-berlin.de and frank.noe@fu-berlin.de

simulation trajectories, enough to reach local equilibrium.[11] (3) We can avoid this problem by preparing local equilibrium distributions in the starting states using biased simulations and then shooting trajectories out of them.[16–19]

Here we qualify and quantify these ideas by systematically analyzing the effect of non-equilibrium starting conditions onto MSM quality, and we suggest effective correction mechanisms. Throughout the manuscript, we use the term "non-equilibrium" to describe the problem that simulations are started from a distribution which is not in global equilibrium, and their simulation time is too short to reach that global equilibrium. Note, however, that the dynamics itself is assumed to possess a unique equilibrium distribution, and if long enough simulations would be run, they would sample from the equilibrium distribution. Briefly, our main results are the following:

1. We provide an expression for the error between unbiased transition probabilities and the expected estimate from many simulations running for multiple discrete time steps, see Sec. II. We find that there is no fundamental advantage of starting simulations in local equilibrium. Rather, the estimation error depends on the discretization, the simulation length, and the lag time. In the limit of long lag times and fine discretization, MSMs are estimated without bias even when non-equilibrium starting points are used. However, for a given discretization the lag time required to practically achieve a small estimation bias might be large.

2. We derive an unbiased MSM estimator that corrects the error due to non-equilibrium starting conditions at short lag times, by exploiting the framework of *observable operator models* (OOMs)—see Sec. III. OOMs are powerful finite-dimensional models that provide unbiased estimates of stationary and kinetic properties of stochastic processes under fairly mild assumptions, see Refs. 20–22. Most importantly, OOMs can be estimated from non-equilibrium simulations[22] and are not limited to a local equilibrium assumption.

3. We utilize the fact that exact relaxation time scales that are not contaminated by the MSM projection error (i.e., quality of the coordinates and the clustering used) can be estimated using the OOM framework. The difference between the unbiased estimate and the uncorrected or corrected MSM estimate is very insightful as it provides an indicator of the quality of the MSM discretization. If this difference is too large, it is suggested to rather improve the coordinate selection or discretization used for MSM construction and re-analyze. Note that while OOMs offer the more general theory, they are not as easy to interpret and their estimation from finite data is not as stable and mature as MSM estimation.

As a technical advance, we provide a meaningful strategy to select the model rank of an OOM which is required in order to obtain practically useful estimates, by using a statistical analysis of singular values of the count matrix (Sec. III D).

Sec. IV demonstrates the usefulness of the OOM framework for two model systems and MD simulation data of alanine dipeptide. We show that accurate estimates of spectral and stationary properties can be obtained from short non-equilibrium simulations, even for short lag times or poor discretizations. We explain how the discretization quality is revealed by the difference between spectral estimates of the MSM and OOM. We also show that the rank selection strategy helps to choose a suitable model rank even for small lag times, when no apparent time scale separation can be utilized.

As an illustration, consider the one-dimensional model system governed by the potential shown in Fig. 1(a), see
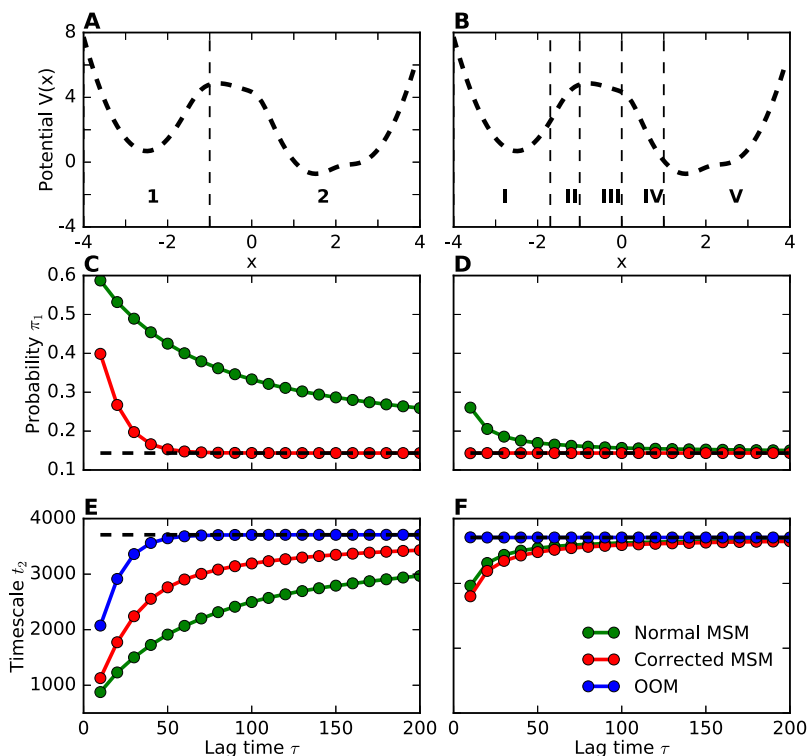

FIG. 1. (a) One-dimensional potential function and discretization into two states. (b) The same potential with a five-state discretization. ((c) and (d)) Estimates for the equilibrium probability of state 1 from the direct MSM (green) and the unbiased MSM (red), reference in black. ((e) and (f)) Estimates for the slowest relaxation time scale t_2 from a direct MSM (green), c.f. Eq. (19), the unbiased MSM (red), c.f. Eqs. (38) and (39), and the spectral OOM estimation (blue), Eqs. (62) and (63). The black dashed line corresponds to the reference value.

Sec. IV A for details. We study the estimation of a Markov model using the two-state discretization indicated in panel (a) of Fig. 1. For various lag times, we investigate the expected transition matrix if 90% of the simulations are started from local equilibrium within state 1, while the other 10% are started from local equilibrium within state 2. Note that we do not use any simulation data here, we only compute expected values over an ensemble of trajectories, with the trajectory length set to 2000 steps, which is shorter than the slowest relaxation time scale.

For short lag times, the standard MSM provides a strongly biased estimate of the equilibrium population of the two wells (Fig. 1(c), green curve). For longer lag times, the MSM converges towards the correct equilibrium population, but the bias only disappears when the lag time approaches the longest relaxation time scale of the system, so if the initial distribution is far from equilibrium this can entail a significant error at practically feasible lag times. In contrast, the corrected MSM estimate proposed in this paper achieves the correct estimate of equilibrium populations even at short lag times (Fig. 1(c), red curve). The standard MSM relaxation time scales are underestimated at short lag times, consistent with previous variational results,[23–25] but they can be improved by using the unbiased MSM estimator proposed here (Fig. 1(e)). The OOM can provide a model-free estimate of the relaxation time scale that is unbiased at a relatively short lag time (Fig. 1(e), blue line). The difference between the OOM and the corrected MSM estimate (blue versus red lines in Fig. 1(e)) is an indicator of the MSM model error due to the state space discretization. Please note that all MSM results in this figure can be dramatically improved if a finer clustering is used. For example, if the five-state partitioning from Fig. 1(b) is used instead, the estimation of stationary properties converges much faster (Fig. 1(d)), and there is hardly a difference between the time scales estimated by a direct and an unbiased MSM (Fig. 1(f)).

II. ANALYSIS OF THE STATE OF THE ART: MSM ESTIMATION FROM SIMULATIONS WITH ARBITRARY STARTING POINTS

A. Molecular dynamics, count matrix, and transition matrix

In this work, we consider the setting described in detail in Ref. 1, that is, an ergodic and reversible Markov process X_t on continuous state space Ω, which possesses a unique stationary distribution π. We denote by $\tau > 0$ the lag time and by

$$p(x, y; \tau) = \mathbb{P}(X_\tau \in dy | X_0 = x) \qquad (1)$$

the conditional transition density function, that is the probability that the process, when located at configuration x at time t, will be found at configuration y at time $t + \tau$. The corresponding transfer operator is denoted by $\mathcal{T}(\tau)$ and is defined by its action on a function of state space u,

$$\mathcal{T}(\tau)u(y) = \int_\Omega p(x, y; \tau) \frac{\pi(x)}{\pi(y)} u(x) \, dx. \qquad (2)$$

Its eigenvalues are called

$$\lambda_m(\tau) = \exp(-\tau/t_m), \qquad (3)$$

where t_m are the implied relaxation time scales. We denote the transfer operator eigenfunctions by ψ_m, $m = 1, \dots$. In particular, we have that $\psi_1 \equiv 1$. If the transfer operator is of rank M at lag time τ, the transition density can be written as

$$p(x, y; \tau) = \sum_{m=1}^M \lambda_m(\tau) \psi_m(x) \pi(y) \psi_m(y). \qquad (4)$$

Note that exact equality in Eq. (4) is an assumption, but often it is satisfied approximately for a large range of lag times τ. Throughout the paper, we will consider decompositions of state space into disjoint sets $S_1, \dots, S_N$, where $\Omega = \bigcup_i S_i$. The indicator function of set S_i is called χ_i. For a simulation of the continuous dynamics which samples positions at discrete time steps, we will denote the position at the kth time step by X_k, $k = 1, \dots, K$, such that K is the total number of time steps in the simulation. We use the symbol $\mathbf{Y}$ as a shorthand notation for an entire simulation. If multiple different simulations need to be distinguished, we will denote them by $\mathbf{Y}_q$, $q = 1, \dots, Q$, i.e., Q is the total number of available simulations.

Most of this work is based on correlations between the discrete sets. For a trajectory as above, we define the empirical histograms and correlations (also called state-to-state time-correlations) as follows:

$$\mathbf{s}(i) := \frac{1}{K - 2\tau} \sum_{k=1}^{K-2\tau} \chi_i(X_k), \qquad (5)$$

$$\mathbf{S}^\tau(i, j) := \frac{1}{K - 2\tau} \sum_{k=1}^{K-2\tau} \chi_i(X_k)\chi_j(X_{k+\tau}), \qquad (6)$$

$$\mathbf{S}_r^{2\tau}(i, j) := \frac{1}{K - 2\tau} \sum_{k=1}^{K-2\tau} \chi_i(X_k)\chi_r(X_{k+\tau})\chi_j(X_{k+2\tau}). \quad (7)$$

Up to the normalization, the matrix $\mathbf{S}^\tau \in \mathbb{R}^{N \times N}$ is a *count matrix* because it simply counts the number of transitions from state S_i to S_j over a time window τ that have occurred in the simulation, while the vector $\mathbf{s} \in \mathbb{R}^N$ counts the total visits to state S_i and corresponds to the ith row sum of $\mathbf{S}^\tau$. For each set S_r, the matrix $\mathbf{S}_r^{2\tau} \in \mathbb{R}^{N \times N}$ is proportional to a two-step count matrix counting subsequent transitions from state S_i to S_r and on to state S_j. At first sight, it may seem confusing that $\mathbf{S}^\tau$ and $\mathbf{s}$ only count transitions and visits up to time $K - 2\tau$, but further below, we will use all three matrices in conjunction which requires estimating all of them over the same part of the data. We will continue to refer to these matrices as a count matrix, count vector, and two-step count matrix in what follows. Also note that in the literature, the count matrix and vector are often denoted by $\mathbf{C}^\tau$, $\mathbf{c}$, but we will use these symbols differently in what follows. Let us note at this point that $\mathbf{s}$, $\mathbf{S}^\tau$, $\mathbf{S}_r^{2\tau}$ can be seen as random variables that map a (stochastic) trajectory $\mathbf{Y}$ of discrete time steps to the values given in Eqs. (5)–(7). To emphasize this dependence, we will also write $\mathbf{s}(\mathbf{Y})$, $\mathbf{S}^\tau(\mathbf{Y})$, $\mathbf{S}_r^{2\tau}(\mathbf{Y})$ if appropriate.

We are concerned with the estimation of a transition probability matrix between the sets S_i of a given discretization of

state space. If the process is in equilibrium, the conditional transition probabilities can be expressed as

$$\mathbf{T}_{Eq}^{\tau}(i,j) = \frac{\mathbb{P}\left(X_t \in S_i,\ X_{t+\tau} \in S_j\right)}{\mathbb{P}(X_t \in S_i)} \tag{8}$$

$$= \frac{\int_{S_i} dx \int_{S_j} dy\ \pi(x)\, p(x,y;\tau)}{\int_{S_i} dx\ \pi(x)} \tag{9}$$

$$= \frac{\mathbf{C}_{Eq}^{\tau}(i,j)}{\pi_i}. \tag{10}$$

Here, we have defined the *equilibrium correlation* between sets S_i and S_j by the nominator of Eq. (9) and denoted it by $\mathbf{C}_{Eq}^{\tau}(i,j)$. Also, we have adopted the usual notation $\pi_i = \int_{S_i} dx\ \pi(x)$ for the equilibrium probabilities of the discrete states. Such a matrix of conditional transition probabilities is called a Markov state model (MSM) or Markov model. It can be used as a simplified model for the dynamics allowing extensive analysis, see Ref. 1.

From a long simulation X_k, $k = 1, \ldots, K$ that samples points from the stationary density π, the matrix $\mathbf{T}_{Eq}^{\tau}$ can be estimated by the formula

$$\mathbf{T}_{Eq}^{\tau}(i,j) \approx \frac{\mathbf{S}^{\tau}(i,j)}{\mathbf{s}(i)}. \tag{11}$$

B. Starting from local equilibrium

In practice, producing simulation data that sample from the global equilibrium density π is often not tractable. One of the strengths of Markov models is the fact that the transition matrix can also be expressed in terms of local equilibrium densities

$$\pi_{S_i}(x) = \frac{1}{\pi_i} \chi_i(x)\pi(x). \tag{12}$$

The density π_{S_i} is the normalized restriction of π to state S_i. A Markov model transition matrix can also be estimated by preparing an ensemble of trajectories in such a way that, within each state, the distribution of starting points equals the local density Eq. (12). These trajectories are simulated for a very short time, and the fraction of trajectories starting in S_i and ending up in S_j provides an estimate for the transition matrix entry $\mathbf{T}_{Eq}^{\tau}(i,j)$.[16,26] To see this, note that in the setting just described, the initial distribution is a convex combination ρ_L of the local densities π_{S_i},

$$\rho_L = \sum_{i=1}^{N} a_i \pi_{S_i}, \tag{13}$$

$$\sum_{i=1}^{N} a_i = 1. \tag{14}$$

Here, a_i is the probability to start in state S_i. Upon replacing π by ρ_L in Eq. (9), it follows that

$$\mathbf{T}_{Eq}^{\tau}(i,j) = \frac{\int_{S_i} dx \int_{S_j} dy\ \rho_L(x)p(x,y;\tau)}{\int_{S_i} dx\ \rho_L(x)}. \tag{15}$$

Only very short trajectories and knowledge of the local densities are needed for the application of this method. However, this method suffers from three major disadvantages: first, the intermediate data points of the simulations cannot be used. Second, estimation of the local densities requires the use of biased sampling methods, which is a significant extra effort and entails additional difficulties. Third, changing the discretization requires redoing the simulations, which is not acceptable if a suitable discretization is not easy to find.

C. Multiple-step estimator

A common way to construct MSMs in practice is by conducting a large set of distributed simulations $\mathbf{Y}_q$, $q = 1, \ldots, Q$ of lengths that are shorter than the largest relaxation time scales of the system but are longer than the lag time τ. For our theoretical investigation, we will assume that each of these trajectories has the same length of K stored simulation steps, but for the estimators we will be deriving later that uniform length is not a requirement, see Appendix C.

The simulations are started from some arbitrary initial distribution at time $k = 1$. The transition probability matrix is estimated by replacing $\mathbf{S}(i,j)$ and $\mathbf{s}(i)$ by their empirical mean values over all simulations $\mathbf{Y}_q$. These are defined by the following equations, where we include the corresponding definition for $\mathbf{S}_r^{2\tau}$ for later use,

$$\overline{\mathbf{s}} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{s}(\mathbf{Y}_q), \tag{16}$$

$$\overline{\mathbf{S}}^{\tau} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{S}^{\tau}(\mathbf{Y}_q), \tag{17}$$

$$\overline{\mathbf{S}}_r^{2\tau} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{S}_r^{2\tau}(\mathbf{Y}_q). \tag{18}$$

In analogy to Eq. (11), the transition matrix is then estimated by

$$\overline{\mathbf{T}}^{\tau}(i,j) = \frac{\overline{\mathbf{S}}^{\tau}(i,j)}{\overline{\mathbf{s}}(i)}. \tag{19}$$

Additional constraints can be incorporated in order to obtain more specific estimators than Eq. (19), such as estimators obeying detailed balance.[1,10,27]

The argument from Sec. II B cannot be transferred directly to a multiple-step estimator like Eqs. (16) and (17): Even if the simulations are started from local equilibrium, this property is lost after the first simulation step, and the resulting estimates are no longer unbiased. A detailed illustration of this phenomenon has been provided by Ref. 1 [Fig. 4], and we repeat it here in Fig. 2. It can be argued that if the discretization is chosen well enough such that the dynamics equilibrates to an approximate local equilibrium within all states over a single time step, the bias can be expected to be very small. This assumption is difficult to check or quantify in practice. In Sec. II D, we analyze the bias introduced by the multiple-step estimator, as well as its dependence on the lag time and simulation length.
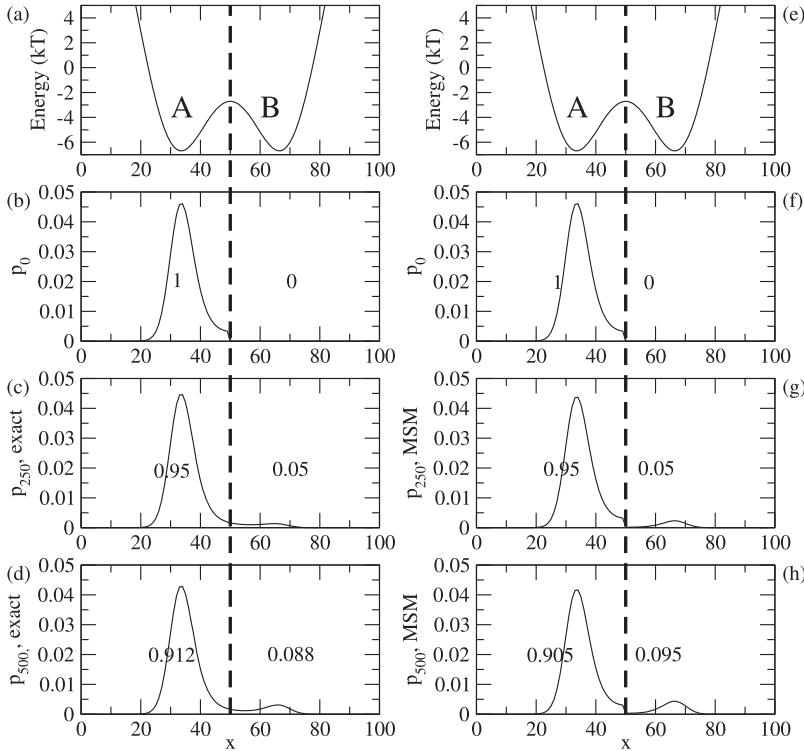
FIG. 2. Loss of local equilibrium property illustrated by comparing the dynamics of the diffusion in a double-well potential (a) and (e) at time steps 0 (b), 250 (c), 500 (d) with the predictions of a Markov model parameterized at lag time $\tau = 250$ at the same times 0 (f), 250 (g), 500 (h). Please refer to the supplementary material of Ref.[1] for details of the system. (b)–(d) show the true distribution of the system (solid black line) and the probabilities associated with the two discrete states left and right of the dashed line. The numbers in (f)–(h) are the discrete state probabilities $p_i(k\tau)$, $i = 1, 2$, $k = 0, 1, 2$, predicted by the Markov model. The solid black lines show the hypothetical density $p_i(k\tau)\pi_{S_i}$ that is inherently assumed when estimating a Markov model by counting transitions over multiple steps. This figure has been re-used with permission from Prinz *et al.*, J. Chem. Phys. **134**, 174105 (2011) [Fig. 4]. Copyright 2011 American Institute of Physics.

D. Estimation error from non-equilibrium simulations

Now we study the effect of using an initial distribution of simulation data that is not in local equilibrium when the transitions are counted. This deviation from local equilibrium could come either from the fact that we start trajectories in an arbitrary initial condition, or that our trajectories exceed the lag time τ such that an initially prepared local equilibrium is lost for all transition counts harvested after the first one (Sec. II C).

Let ρ denote the *empirical* distribution sampled by the simulations. We need to study the error between the equilibrium transition matrix $\mathbf{T}_{Eq}^{\tau}$ and the asymptotic limit of Eq. (19). To this end, we study the asymptotic limits of $\overline{\mathbf{S}}^{\tau}(i,j)$ and $\bar{\mathbf{s}}(i)$ in the limit of infinitely many simulations, $Q \to \infty$, but each having finite lengths,

$$\mathbf{C}_{\rho}^{\tau}(i,j) := \mathbb{E}\left(\mathbf{S}^{\tau}(i,j)\right), \tag{20}$$

$$\mathbf{c}_{\rho}(i) := \mathbb{E}\left(\mathbf{s}(i)\right), \tag{21}$$

$$\mathbf{T}_{\rho}^{\tau}(i,j) := \frac{\mathbf{C}_{\rho}^{\tau}(i,j)}{\mathbf{c}_{\rho}(i)}. \tag{22}$$

Thus, we use the symbols $\mathbf{C}_{\rho}^{\tau}$, $\mathbf{c}_{\rho}$ for the *expected* count matrix and vector of total counts associated with the empirical distribution ρ. Using the spectral decomposition, Eq. (4), the expected count matrix can be expressed in terms of the spectral components of the dynamics,

$$\mathbf{C}_{\rho}^{\tau}(i,j) = \sum_{m=1}^{M} \lambda_m(\tau) \langle \chi_i, \psi_m \rangle_{\rho} \langle \chi_j, \psi_m \rangle_{\pi}, \tag{23}$$

$$\langle \chi_i, \psi_m \rangle_{\rho} = \int_{\Omega} \mathrm{d}x \, \chi_i(x)\psi_m(x)\rho(x), \tag{24}$$

$$\langle \chi_i, \psi_m \rangle_{\pi} = \int_{\Omega} \mathrm{d}x \, \chi_i(x)\psi_m(x)\pi(x). \tag{25}$$

In matrix form, Eq. (23) can be written as

$$\mathbf{C}_{\rho}^{\tau} = \mathbf{Q}_{\rho}\mathbf{\Lambda}(\tau)\mathbf{Q}_{\pi}^{T}, \tag{26}$$

$$\mathbf{Q}_{\rho}(i,m) = \langle \chi_i, \psi_m \rangle_{\rho}, \tag{27}$$

$$\mathbf{Q}_{\pi}(j,m) = \langle \chi_j, \psi_m \rangle_{\pi}. \tag{28}$$

These matrices contain the MSM projections of the true eigenfunctions, i.e., their approximations by step functions, which are extensively discussed in Refs. 1 and 14. Let us emphasize that Eq. (23) also holds for arbitrary basis functions, i.e., χ_i is not required to be a basis of indicator functions. Thus, it is the most general expression for a correlation matrix from Markovian dynamics.

Summation over j shows that

$$\mathbf{c}_{\rho}(i) = \langle \chi_i \rangle_{\rho}. \tag{29}$$

It follows from Eq. (23) that the spectral expansion of $\mathbf{C}_{Eq}^{\tau}$ is given by

$$\mathbf{C}_{Eq}^{\tau}(i,j) = \sum_{m=1}^{M} \lambda_m(\tau) \langle \chi_i, \psi_m \rangle_{\pi} \langle \chi_j, \psi_m \rangle_{\pi}, \tag{30}$$

using the fact that for trajectories started from global equilibrium we have $\rho = \pi$. Combining Eqs. (23), (29), and (30), we obtain an expression for the estimation error $\mathbf{E}^{\tau} := \mathbf{T}_{\rho}^{\tau} - \mathbf{T}_{Eq}^{\tau}$,

$$\mathbf{E}^\tau(i,j) = \frac{\mathbf{C}^\tau_\rho(i,j)}{\mathbf{c}_\rho(i)} - \frac{\mathbf{C}^\tau_{Eq}(i,j)}{\pi_i} \qquad (31)$$

$$= \sum_{m=2}^{M} \lambda_m(\tau)\langle\chi_j,\psi_m\rangle_\pi \left[\frac{\langle\chi_i,\psi_m\rangle_\rho}{\langle\chi_i\rangle_\rho} - \frac{\langle\chi_i,\psi_m\rangle_\pi}{\langle\chi_i\rangle_\pi} \right] \quad (32)$$

$$= \sum_{m=2}^{M} \lambda_m(\tau)\langle\chi_j,\psi_m\rangle_\pi \left[\frac{\langle\chi_i,\psi_m - q_{im}\psi_1\rangle_\rho}{\langle\chi_i\rangle_\rho} \right], \qquad (33)$$

where $q_{im} = \frac{\langle\chi_i,\psi_m\rangle_\pi}{\langle\chi_i\rangle_\pi}$, and we were able to drop the $m = 1$ terms on both sides as they are equal. Inspecting this expression leads to a number of insights that are practically important for analyzing simulation data with MSMs:

1. *MSM estimation from long trajectories*: In the limit that our trajectories are longer than the time scale of the slowest process, the empirical distribution ρ converges to the equilibrium distribution π, and the bias becomes zero. This offers an explanation why MSMs built from ultra-long simulations[28,29] are quite well-behaved and have been extensively used for benchmarking and method validation.

2. *MSM estimation from short trajectories*: Even if the trajectories are not long enough to reach global equilibrium, because of Eq. (3), the bias decays multi-exponentially with the lag time τ. This is an important insight because MSMs are in practice constructed in the limit of long enough lag times in which the time scale estimates converge,[1,5] and the above equation shows that this limit is meaningful as it approaches an unbiased estimate.

3. *Dependence of bias on the discretization error*: The above formula reflects the well-known insight that Markov models are free of bias if the discretization perfectly approximates the dominant eigenfunctions, meaning that the eigenfunctions are constant on the states S_i.[1,5]

4. *Consequences for adaptive sampling*: Previous adaptive sampling approaches have suggested preparing an initial local equilibrium distribution in order to shoot trajectories out of selected states.[16] The above analysis shows that this strategy is effective if we only count a single transition out of the state but is ineffective when longer trajectories are shot. In the latter case, it is simpler to

ignore the initial distribution and to reduce the effect of bias by extending the lag time τ, see again Fig. 1 and also the next example.

E. Example

Before proceeding, we illustrate these findings by revisiting the one-dimensional model system presented in the Introduction. We study the same two different discretizations, the two-state model from panel (a) of Fig. 3 and the five-state discretization shown in Fig. 3(b). Again, simulations are initiated from local equilibrium in states 1 and 2 of the coarse discretization, with $a_1 = 0.9$, $a_2 = 0.1$. We study the expected estimate of the equilibrium probability of state 1, which equals the equilibrium probability of states I and II for the finer state definition. Panels (c) and (d) of Fig. 3 show the respective estimates for the coarse and fine discretization as a function of the lag time, for simulation lengths $K = 1000, 2000, 5000, 10\,000, 50\,000$. Indeed, the estimates improve if the lag time is increased, if the simulation length is increased, or if the discretization is improved. From the coarse partitioning example, we conclude that relaxation to global equilibrium can be required in order to obtain unbiased estimates from simulations initiated out of local equilibrium.

III. CORRECTION OF ESTIMATION BIAS USING OBSERVABLE OPERATOR MODELS

In this section, we show how to go beyond just using a longer lag time τ and suggest correction mechanisms to obtain the correct equilibrium transition matrix $\mathbf{T}^\tau_{Eq}$ (Eqs. (8)–(10)) from an ensemble of short simulations. This can be accomplished regardless of the starting distribution being in global equilibrium, in local equilibrium, or far from any equilibrium.

As discussed above, limitations of MSMs include the assumption of Markovianity, sensitivity to projection error, and sensitivity to the distribution of trajectory starting points. All of these limitations can be overcome by realizing that molecular dynamics that is observed in a chosen set of variables, reaction coordinates, or order parameters at a certain lag time τ can be *exactly* described by projected Markov models
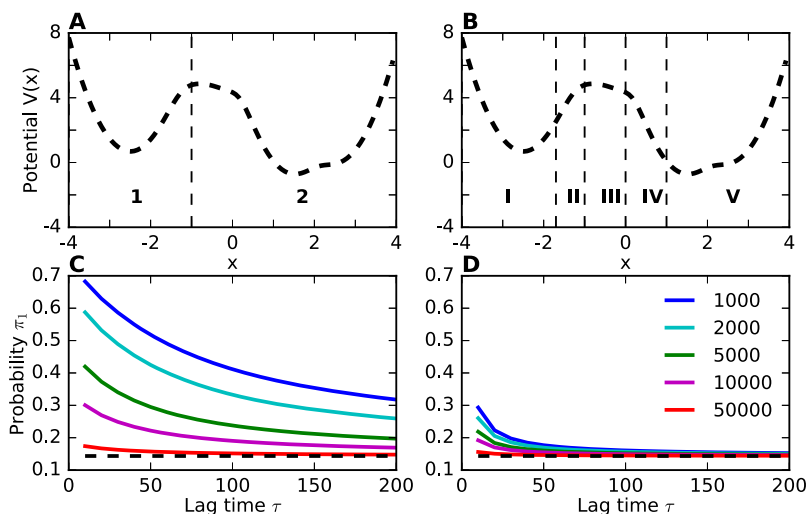

FIG. 3. ((a) and (b)) One-dimensional potential function with two different discretizations into two states and five states, respectively. (c) Expected estimate of the equilibrium probability of state 1 as a function of the lag time, for simulation lengths $K = 1000, 2000, 5000, 10\,000, 50\,000$, and using the discretization from panel (a). The simulations are initiated in local equilibrium in both states 1 and 2 but predominantly in state 1 ($a_1 = 0.9$, $a_2 = 0.1$). (d) The same for the five-state discretization from panel (b).

(PMMs).[30] This insight allows us to employ estimators that are not affected by the MSM limitations, such as hidden Markov models (HMMs)[30] or observable operator models (OOMs),[20–22] that operate on the discretized state space.

Here, we employ OOMs in order to get improved MSM estimators that are not subject to the bias caused by a non-equilibrium distribution of the trajectories used. In a nutshell, OOMs are spectral estimators able to provide unbiased estimates of stationary and dynamical quantities for dynamical systems that can be well described by a finite number of dynamical components. Here we only summarize a few aspects of OOMs that are relevant to the present paper and present an algorithm that can be used to estimate MSMs without bias from the initial trajectory distribution. To fully understand the theoretical background and derivation, please refer to Refs. 20–22.

A. Observable operator models

Observable operator models (OOMs) provide a framework that completely captures the dynamics of a stochastic dynamical system by a finite-dimensional algebraic system if only a finite number M of relaxation processes contribute in Eq. (4), see Refs. 20 and 21. For molecular dynamics, this property is achieved if we observe and model the dynamics at a finite lag time τ. The *full-state observable operator* $\boldsymbol{\Xi}_\Omega$ is an $M \times M$ matrix which contains the scalar products between the eigenfunctions,

$$\boldsymbol{\Xi}_\Omega(m, m') = \lambda_m(\tau) \int_\Omega dx\, \psi_m(x) \psi_{m'}(x) \pi(x). \quad (34)$$

In statistical terms, $\boldsymbol{\Xi}_\Omega$ is the expectation value of the covariance matrix between eigenfunctions. As eigenfunctions are orthogonal with respect to the equilibrium distribution π, or in other words, statistically uncorrelated, $\boldsymbol{\Xi}_\Omega$ is just a diagonal matrix of the eigenvalues,

$$\boldsymbol{\Xi}_\Omega = \boldsymbol{\Lambda}. \quad (35)$$

If we do not integrate over the full state space Ω in Eq. (34), but only over a subset $A \subset \Omega$, we can define a matrix $\boldsymbol{\Xi}_A$ of size $M \times M$, called the *set-observable operator* for set A. All set-observable operators and two vectors $\boldsymbol{\omega}, \boldsymbol{\sigma} \in \mathbb{R}^M$ are the key ingredients of OOM theory. The vectors $\boldsymbol{\omega}, \boldsymbol{\sigma}$ equal the first canonical unit vector $\mathbf{e}_1$, i.e., $\boldsymbol{\omega} = \boldsymbol{\sigma} = \mathbf{e}_1 = (1, 0, \ldots, 0)^T$, and they are called *information state* and *evaluator*, respectively. If the finite-rank assumption Eq. (4) holds, these components form an algebraic system that allows to compute equilibrium probabilities of finite observation sequences. Let $A_1, \ldots, A_l$ be arbitrary subsets of Ω that do not need to form a partition of the state space. If Eq. (4) is satisfied, we can compute the probability that a trajectory in equilibrium visits set A_1 at time τ, set A_2 at time 2τ, ..., and set A_l at time $l\tau$ by the following matrix-vector product:

$$\mathbb{P}(X_\tau \in A_1, X_{2\tau} \in A_2, \ldots, X_{l\tau} \in A_l) = \boldsymbol{\omega}^T \boldsymbol{\Xi}_{A_1} \ldots \boldsymbol{\Xi}_{A_l} \boldsymbol{\sigma}. \quad (36)$$

The proof can be found in Ref. 21, we also repeat it in Appendix B. Note that, in case that $A_1, \ldots, A_l$ form a partition of state space, the probability of such an observation sequence cannot be computed from a Markov model transition matrix between the sets $A_1, \ldots, A_l$, unless the dynamics is

Markovian on these sets. This clearly distinguishes an OOM from a Markov model: An OOM can correctly describe arbitrary projected dynamics as long as Eq. (4) holds.

As a Markov process is determined entirely by finite observation probabilities like Eq. (36), it follows that we can compute several key equilibrium, kinetic, and mechanistic quantities in an unbiased fashion if we can somehow estimate the OOM components. For a fixed decomposition of state space into sets S_r, $r = 1, \ldots, N$ as before, let us denote the set-observable operators of sets S_r by $\boldsymbol{\Xi}_r$, which implies that

$$\boldsymbol{\Xi}_\Omega = \sum_{r=1}^N \boldsymbol{\Xi}_r. \quad (37)$$

It follows from Eq. (36) that we can compute the unbiased equilibrium correlation matrix and the stationary probabilities by the formulas

$$\mathbf{C}_{Eq}^\tau(i, j) = \boldsymbol{\omega}^T \boldsymbol{\Xi}_i \boldsymbol{\Xi}_j \boldsymbol{\sigma}, \quad (38)$$

$$\pi_i = \boldsymbol{\omega}^T \boldsymbol{\Xi}_i \boldsymbol{\sigma}. \quad (39)$$

In practice we cannot directly estimate $\boldsymbol{\Xi}_r$ but only a *similar* operator $\hat{\boldsymbol{\Xi}}_r$. However, it follows directly from Eqs. (38) and (39) that if an unknown similarity transform $\mathbf{R} \in \mathbb{R}^{M \times M}$ affects all OOM quantities via

$$\hat{\boldsymbol{\Xi}}_r = \mathbf{R} \boldsymbol{\Xi}_r \mathbf{R}^{-1}, \quad (40)$$

$$\hat{\boldsymbol{\omega}}^T = \boldsymbol{\omega}^T \mathbf{R}^{-1}, \quad (41)$$

$$\hat{\boldsymbol{\sigma}} = \mathbf{R} \boldsymbol{\sigma}, \quad (42)$$

then Eqs. (38) and (39) remain exactly valid using $\hat{\boldsymbol{\omega}}$, $\hat{\boldsymbol{\Xi}}_r$, $\hat{\boldsymbol{\sigma}}$. In other words, all OOMs that can be constructed by choosing some transformation matrix $\mathbf{R}$ form a family of equivalent OOMs. A specific member of this family *can* be estimated directly from simulation data, and thus we can use it in order to obtain unbiased estimates of Eqs. 38 and 39 even from a large ensemble of trajectories that do not need to sample from global equilibrium. It has been shown in Ref. 22 that Eqs. (47)–(50) in Sec. III B indeed provide the components of an equivalent OOM, i.e., there is an invertible matrix $\mathbf{R}$ such that Eqs. (40)–(42) are satisfied in the absence of statistical noise.

B. Unbiased estimation of Markov state models

To construct an *exact* unbiased estimator, we need three ingredients: (i) the expectation values of the empirical count matrix $\mathbf{C}_\rho^\tau$, (ii) the vector of total counts $\mathbf{c}_\rho$ from Eqs. (20) and (21), and additionally (iii) the two-step count matrices,

$$\mathbf{C}_{\rho,r}^{2\tau} := \mathbb{E}\left(\mathbf{S}_r^{2\tau}\right). \quad (43)$$

As a reminder, expectation values here denote the expectation over a trajectory ensemble sampling from the empirical (non-equilibrium) distribution ρ. In practice, only finitely many simulations are available, and we thus replace $\mathbf{c}_\rho$, $\mathbf{C}_\rho^\tau$ and $\mathbf{C}_{\rho,r}^{2\tau}$ by count vectors and matrices $\bar{\mathbf{s}}$, $\overline{\mathbf{S}}^\tau$ and $\overline{\mathbf{S}}_r^{2\tau}$ (Eqs. (16)–(18)), which are asymptotically unbiased estimators. The unbiased estimation algorithm can be summarized as follows:

1. Obtain the empirical mean $\bar{\mathbf{s}}$, count matrix $\overline{\mathbf{S}}^\tau$, and two-step count matrices $\overline{\mathbf{S}}_r^{2\tau}$ from simulation data using Eqs (16)–(18).

2. Decompose the count matrix $\overline{\mathbf{S}}^\tau$ by singular value decomposition (SVD),

$$\overline{\mathbf{S}}^\tau = \mathbf{V}\Sigma\mathbf{W}^T, \tag{44}$$

and compute weighted projections onto the leading M left and right singular vectors by

$$\mathbf{F}_1 = \mathbf{V}_M \Sigma_M^{-1/2}, \tag{45}$$

$$\mathbf{F}_2 = \mathbf{W}_M \Sigma_M^{-1/2}. \tag{46}$$

We have used the symbols $\mathbf{V}_M$, $\mathbf{W}_M$, Σ_M to denote the restriction of these matrices to their first M columns.

3. Use $\mathbf{F}_1$, $\mathbf{F}_2$ to obtain the set-observable operators $\hat{\boldsymbol{\Xi}}_r$ and the evaluation state vector $\hat{\boldsymbol{\sigma}}$ of an equivalent OOM via

$$\hat{\boldsymbol{\Xi}}_r = \mathbf{F}_1^T \overline{\mathbf{S}}_r^{2\tau} \mathbf{F}_2, \tag{47}$$

$$\hat{\boldsymbol{\sigma}} = \mathbf{F}_1^T \overline{\mathbf{s}}. \tag{48}$$

Compute the full-state observable operator $\hat{\boldsymbol{\Xi}}_\Omega = \sum_{r=1}^N \hat{\boldsymbol{\Xi}}_r$ and obtain the information state vector $\hat{\boldsymbol{\omega}}$ as the solution to the eigenvalue problem,

$$\hat{\boldsymbol{\omega}}^T \hat{\boldsymbol{\Xi}}_\Omega = \hat{\boldsymbol{\omega}}^T, \tag{49}$$

$$\hat{\boldsymbol{\omega}}^T \hat{\boldsymbol{\sigma}} = 1. \tag{50}$$

The normalization Eq. (50) can be achieved by dividing the arbitrarily scaled solution $\hat{\boldsymbol{\omega}}^T$ by $\hat{\boldsymbol{\omega}}^T \hat{\boldsymbol{\sigma}}$.

4. Compute the unbiased equilibrium correlation matrix and unbiased equilibrium distribution by

$$\mathbf{C}_{Eq}^\tau(i,j) = \hat{\boldsymbol{\omega}}^T \hat{\boldsymbol{\Xi}}_i \hat{\boldsymbol{\Xi}}_j \hat{\boldsymbol{\sigma}}, \tag{51}$$

$$\pi_i = \hat{\boldsymbol{\omega}}^T \hat{\boldsymbol{\Xi}}_i \hat{\boldsymbol{\sigma}} \tag{52}$$

$$= \sum_{j=1}^N \mathbf{C}_{Eq}^\tau(i,j) \tag{53}$$

and then obtain the unbiased MSM transition matrix $\mathbf{T}_{Eq}^\tau$ either using the nonreversible estimator

$$\mathbf{T}_{Eq}^\tau(i,j) = \frac{\mathbf{C}_{Eq}^\tau(i,j)}{\pi_i}, \tag{54}$$

or the reversible estimator

$$\mathbf{T}_{Eq}^\tau(i,j) = \frac{\mathbf{C}_{Eq}^\tau(i,j) + \mathbf{C}_{Eq}^\tau(j,i)}{\sum_{j=1}^N \mathbf{C}_{Eq}^\tau(i,j) + \sum_{j=1}^N \mathbf{C}_{Eq}^\tau(j,i)}. \tag{55}$$

Let us briefly comment on the central idea behind this algorithm, which is the estimation of an equivalent OOM in the third step, particularly in Eq. (47). Using the path probability formula Eq. (36), it can be shown that the expected two-step count matrix is given by

$$\mathbf{C}_{\rho,r}^{2\tau} = \mathbf{Q}_\rho \boldsymbol{\Xi}_r \boldsymbol{\Lambda}(\tau) \mathbf{Q}_\pi^T, \tag{56}$$

where the matrices $\mathbf{Q}_\rho$, $\mathbf{Q}_\pi$ are the same as in Eqs. (27) and (28). Thus, by the intermediate step, the set-observable operator is introduced into the decomposition of the two-step count matrix. Now, the idea is to find two matrices $\mathbf{F}_1$, $\mathbf{F}_2 \in \mathbb{R}^{N \times M}$, such that $\mathbf{R}_1 := \mathbf{F}_1^T \mathbf{Q}_\rho$ and $\mathbf{R}_2 := \boldsymbol{\Lambda}(\tau) \mathbf{Q}_\pi^T \mathbf{F}_2$ are inverse to each other, because this implies that

$$\mathbf{F}_1^T \mathbf{C}_{\rho,r}^{2\tau} \mathbf{F}_2 = \mathbf{R}_1 \boldsymbol{\Xi}_r \mathbf{R}_2 \tag{57}$$

$$= \mathbf{R} \boldsymbol{\Xi}_r \mathbf{R}^{-1} \tag{58}$$

is the rth component of an equivalent OOM. The properties of SVD and the decomposition Eq. (26) guarantee that the choice of $\mathbf{F}_1$, $\mathbf{F}_2$ in the second step above achieves this goal,

$$\mathbf{Id} = \mathbf{F}_1^T \mathbf{C}_\rho^\tau \mathbf{F}_2 \tag{59}$$

$$= \left(\mathbf{F}_1^T \mathbf{Q}_\rho\right)\left(\boldsymbol{\Lambda}(\tau)\mathbf{Q}_\pi^T \mathbf{F}_2\right) \tag{60}$$

$$= \mathbf{R}_1 \mathbf{R}_2. \tag{61}$$

Similar arguments can be used to justify the equations for $\boldsymbol{\omega}$, $\boldsymbol{\sigma}$. We also note that different choices of $\mathbf{F}_1$, $\mathbf{F}_2$ in step 2 are possible. For detailed explanations and proofs, please refer to the previous publications.[20–22]

C. Recovery of exact relaxation time scales

A remarkable by-product of the procedure described above is that the transformed full-state two-step count matrix $\hat{\boldsymbol{\Xi}}_\Omega$ is similar to a diagonal matrix of the system eigenvalues $\lambda_m(\tau)$ *without any MSM projection error*. This has been shown for equilibrium data in Ref. 31 and also applies to non-equilibrium data,[21]

$$\hat{\boldsymbol{\Xi}}_\Omega = \mathbf{R}\boldsymbol{\Xi}_\Omega \mathbf{R}^{-1} \tag{62}$$

$$= \mathbf{R}\boldsymbol{\Lambda}(\tau)\mathbf{R}^{-1}. \tag{63}$$

Thus, diagonalization of $\hat{\boldsymbol{\Xi}}_\Omega$ provides an estimate of the leading system eigenvalues, and consequently also of the relaxation rates or time scales, that is not distorted by the fact that we coarse-grain the dynamics to a Markov chain between coarse sets in state space. These eigenvalue and time scale estimates are only subject to statistical error but not to any MSM model error. It is impossible to directly build an MSM that produces these time scales—when an MSM is desired, the time scales can only be approximated, and they will only be correct in the limit of long lag times and good discretization.

However, the fact that we can get a model-free estimate of the eigenvalues and relaxation time scales can be used to assess the discretization quality: According to the variational principle of conformation dynamics,[24] the exact system eigenvalues provide an upper bound to the eigenvalues of the equilibrium transition matrix $\mathbf{T}_{Eq}^\tau$. By comparing the eigenvalues of $\mathbf{T}_{Eq}^\tau$ to those from Eqs. (62) and (63), the MSM discretization error theoretically studied in Refs. 1, 14, and 23 can be practically quantified.

D. Selection of model rank

The above method is theoretically guaranteed to work whenever the number of MSM states N is at least equal to the number M of relaxation processes in Eq. (4), and the count matrix $\mathbf{C}_\rho^\tau$ is of rank M. In the absence of statistical noise, the model rank M can then be determined by the number of non-zero singular values of $\mathbf{C}_\rho^\tau$. For finite data, the numerical rank of $\overline{\mathbf{S}}^\tau$ is not necessarily equal to M, as the singular values can be perturbed by noise. Classical matrix perturbation theory predicts that small singular values will be particularly affected by noise, see, e.g., Ref. 32, and also Fig. 4(a). Including noisy and small singular values can severely affect the accuracy of the method, most likely due to the presence of the matrix of inverse singular values in Eqs. (45) and (46). Also, we expect
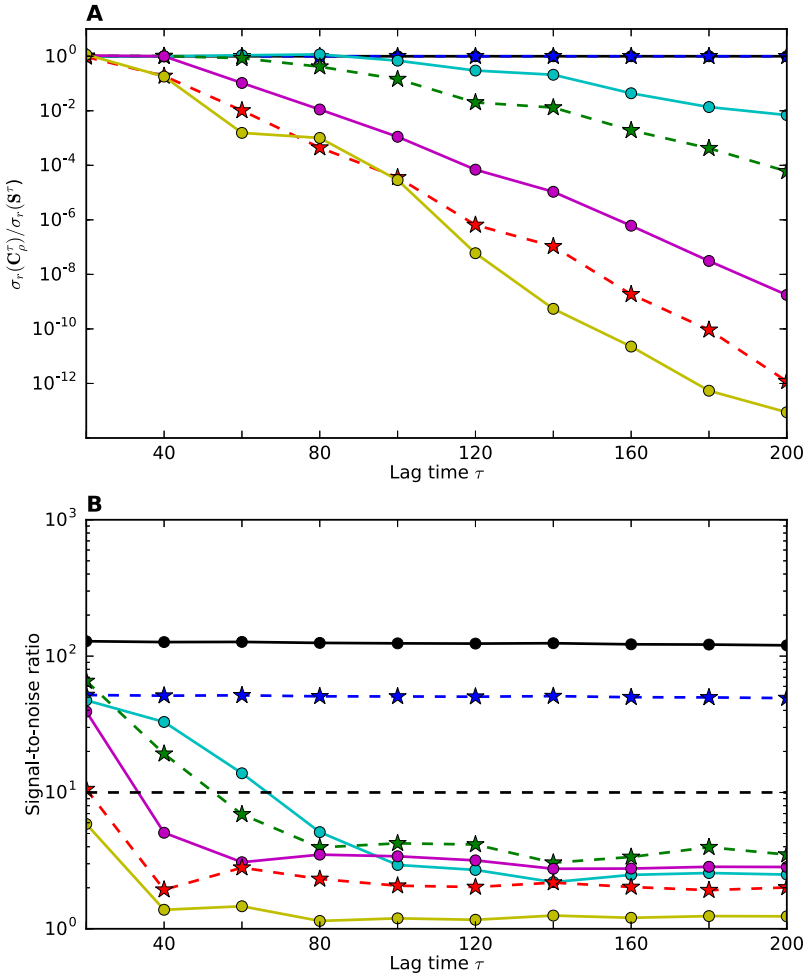
FIG. 4. Analysis of statistical uncertainties for singular values of the count matrix. We use the one-dimensional model system and seven-state discretization as in Sec. IV A, the sample consists of $Q = 5000$ trajectories of length $K = 2000$. (a) For each of the seven singular values (distinguished in descending order by the colors black, blue, cyan, green, magenta, red, and yellow), we show the ratio of the true singular value $\sigma_r(\mathbf{C}_\rho^\tau)$, $r = 1, \ldots, 7$ of the expected count matrix $\mathbf{C}_\rho^\tau$ to the corresponding singular value $\sigma_r(\overline{\mathbf{S}}^\tau)$ of the empirical count matrix $\overline{\mathbf{S}}^\tau$, as a function of the lag time. As the small singular values decay quickly with the lag time, they are dominated by the noise even for small lag times. Including these noisy singular values would ruin the results. (b) Ratio between mean value and uncertainty (signal-to-noise ratio) from the bootstrapping for the seven singular values as a function of the lag time. The thin black dashed line indicates the cutoff we have used in applications. Only singular values above this line are included in the estimation, the number of points above this line corresponds to the OOM model rank, see Fig. 5(h).

small singular values to have little impact on the dominant spectral and stationary properties of the final OOM, but this will be backed up by further theoretical investigation.

Consequently, it seems appropriate to cut off small and statistically unreliable singular values and select a smaller model rank $\hat{M} < M$ in Eqs. (45) and (46). In order to determine the uncertainties of the singular values, we use the bootstrapping procedure, and we discard all singular values with a signal-to-noise ratio of less than 10. This has proven to be a useful choice in all applications presented further below. Fig. 4(b) illustrates this procedure for a simple model system.

E. Software, algorithmic details, and analysis of computational effort

We close the Section III of this paper by pointing out a few more details of practical importance. First, while it was convenient for the theoretical analysis to assume that all trajectories sample the same number of simulation steps K, this is not required (see Appendix C). Moreover, we also argue in Appendix C that all normalizations in Eqs. (5)–(7) and (16)–(18) can be dropped in practice. All of the matrices $\bar{\mathbf{s}}, \overline{\mathbf{S}}^\tau, \overline{\mathbf{S}}_r^{2\tau}$ used in the estimation algorithm can be replaced by integer valued matrices that simply count the number of visits, transitions, and two-step transitions.

Second, we have suggested using the bootstrapping procedure in order to estimate uncertainties for the singular values

of the count matrix. One way to realize this is to re-draw trajectories with replacement from the set of all available simulations and to re-estimate the count matrix from this modified set of simulations. As individual simulations are statistically independent, this procedure is theoretically justified and can also be used to estimate uncertainties of further derived quantities, like time scales and stationary probabilities. We used the trajectory-based bootstrapping in all examples shown below. However, if only a small number of rather long simulations is available, it may be more practical to re-draw individual transitions from the set of all available transitions in the data set. Let T denote the total number of data points, which equals $T = KQ$ for the uniform trajectory length, and Eq. (C1) otherwise. If the transitions were statistically independent, one could simply re-sample T transition pairs from the set of all N^2 possible pairs, where the probability of drawing the pair (i, j) is given by $\overline{\mathbf{S}}^\tau(i, j)$. In fact, transitions are not statistically independent. Therefore, we suggest to replace the count matrix $\overline{\mathbf{S}}^\tau$ by the effective count matrix described in Ref. 33, but it should be noted that this procedure relies on several approximations and must be improved in the future.

Third, we present an overview of the computational cost of each step in the estimation algorithm in Table I, assuming that dense matrix algebra is used in every step. It is expressed in terms of the total number of data points T, the number of

TABLE I. Analysis of computational effort required by the OOM-based estimation algorithm, if all operations are performed in dense matrix algebra.

Operation	Cost
Count matrix estimation	$\propto T$
Bootstrapping	$\propto n_b T N^3$
SVD of $\overline{\mathbf{S}}^\tau$	$\propto N^3$
Computation of OOM components	$\hat{\boldsymbol{\sigma}} : MN + N^2$
	$\hat{\boldsymbol{\Xi}} : N\left(N^2 M + N M^2\right)$
	$\hat{\boldsymbol{\omega}} : \propto M^3 + N M^2$
Transition matrix $\mathbf{T}_{Eq}^\tau$	$N\left(2M^2 + M\right)$

MSM states N, the OOM model rank M, and the number of bootstrapping samples n_b.

The first step requires an effort which is linear in the data size and can be performed efficiently. In most cases, we can also assume the count matrices $\overline{\mathbf{S}}^\tau$, $\overline{\mathbf{S}}_r^{2\tau}$ to be sparse, and the model rank M to be small. In this case, the cubic term appearing for the calculation of $\hat{\boldsymbol{\Xi}}$ becomes quadratic, while the contributions of the model rank are small. The only real bottleneck is the singular value decomposition of $\overline{\mathbf{S}}^\tau$, accounting for the factor N^3 in the second and third steps. As we generally require all singular values of the count matrix, this step must be performed using dense matrix algebra, which can be time-consuming. Future research may provide a method that only requires the computation of the leading singular values, thus allowing for sparse algebra to be employed.

Lastly, we note that the MSM correction method described in Sec. III B is available as part of the PyEMMA package,[34] version 2.3 or later, see http://pyemma.org.

IV. EXAMPLES

For each of the following examples, we use the trajectory-based bootstrapping strategy to determine the OOM model rank. Mean values and standard errors for the singular values are estimated from $n_b = 10\,000$ re-samplings, singular values with a signal-to-noise ratio of at least 10.0 are accepted. We also generate error estimates for all quantities derived from the OOM-based Markov model by trajectory bootstrapping, using 1000 re-samplings. In addition, we compute a conventional Markov model without OOM-based correction as a comparison.

A. One-dimensional toy potential

As the first example, we study in more detail the one-dimensional system used in the Introduction. The system is defined by the double-well potential function shown in Fig. 5(a). The dynamics here is a finite state space Markov chain with 100 microstates distributed along the x-axis, where transitions can occur between neighboring states based on a Metropolis criterion. The system is kinetically two-state, as the slowest relaxation time scale of the system, corresponding to the transition process between the two wells, is $t_2 = 3708$ steps and clearly dominates all others (Fig. 5(b)).

We investigate the estimation of a seven-state Markov model ($N = 7$) using the discretization indicated by dashed lines in Fig. 5(a). Using seven states instead of two accelerates the convergence of OOM estimates. Still, the seven-state discretization is a poor one—note that state 4 contains large parts of the transition region as well as parts of the right minimum. This choice was made deliberately in order to test the robustness of our method with respect to poor MSM clusterings. We produced two different data sets, each comprising $Q = 5000$ simulations. The first set contains short simulations of length $K = 250$, while the simulations of the second set are $K = 2000$ steps long. For the analysis of the smaller data set, we can use lag times up to $\tau = 30$, while we can go up to $\tau = 200$ for the larger data set. Panels (c), (e), and (g) of Fig. 5 display the results for the short simulations, while the corresponding results for the larger data set are shown in panels (d), (f), and (h). All simulations were initiated from a non-equilibrium starting distribution, where the probabilities to start in each of the seven states are given by the vector

$$\rho_1 = \left[\ 0.3\ 0.3\ 0.3\ 0\ 0.05\ 0.05\ 0\ \right], \tag{64}$$

that is, 90% of the simulations were started in the left three states, while only 10% were initialized in the deeper minimum on the right. Within each state, the actual microstate was selected from a uniform distribution.

Figs. 5(c) and 5(d) compare estimates of stationary probabilities from direct MSMs based on Eq. (19) and corrected MSMs with the transition matrix given by Eq. (55). Due to the non-equilibrium initial distribution, the simulations visit the left minimum much more frequently than a simulation in equilibrium would do. While the MSM estimates of the stationary distribution converge to the true equilibrium distribution at long lag times, they are surprisingly inaccurate at short times, where the effect of the non-equilibrium starting distribution still has a strong effect. Even at the largest lag time $\tau = 200$, the bias is still visible. In contrast, the corrected MSM provides an excellent and stable estimate at lag times of 15 steps or longer.

In Figs. 5(e) and 5(f), we compare estimates of the slowest implied relaxation time scale t_2 from three different estimators: a direct Markov model based on Eq. (19), the corrected Markov model based on Eq. (55), and the OOM-based spectral estimation, Eqs. (62) and (63). First, we notice that the direct and corrected MSMs provide different estimates because of the combination of non-equilibrium starting points and the poor discretization quality. The corrected MSM time scales converge faster to the true time scales than the uncorrected ones. Second, the OOM-based direct estimation of relaxation time scales by Eq. (63) provides accurate results already at lag time $\tau = 15$, which is a regime where the number of relevant relaxation processes cannot be easily determined by a time scale separation, see again panel (b) of Fig. 5. The OOM time scale estimates become very accurate for larger lag times if more data can be used. Third, the large deviation between the corrected MSM and the OOM time scales are indicative of the poor discretization quality employed here.

Finally, in Figs. 5(g) and 5(h) we show the model rank selected by the bootstrapping procedure as a function of the lag time. We can observe how our criterion based on statistical
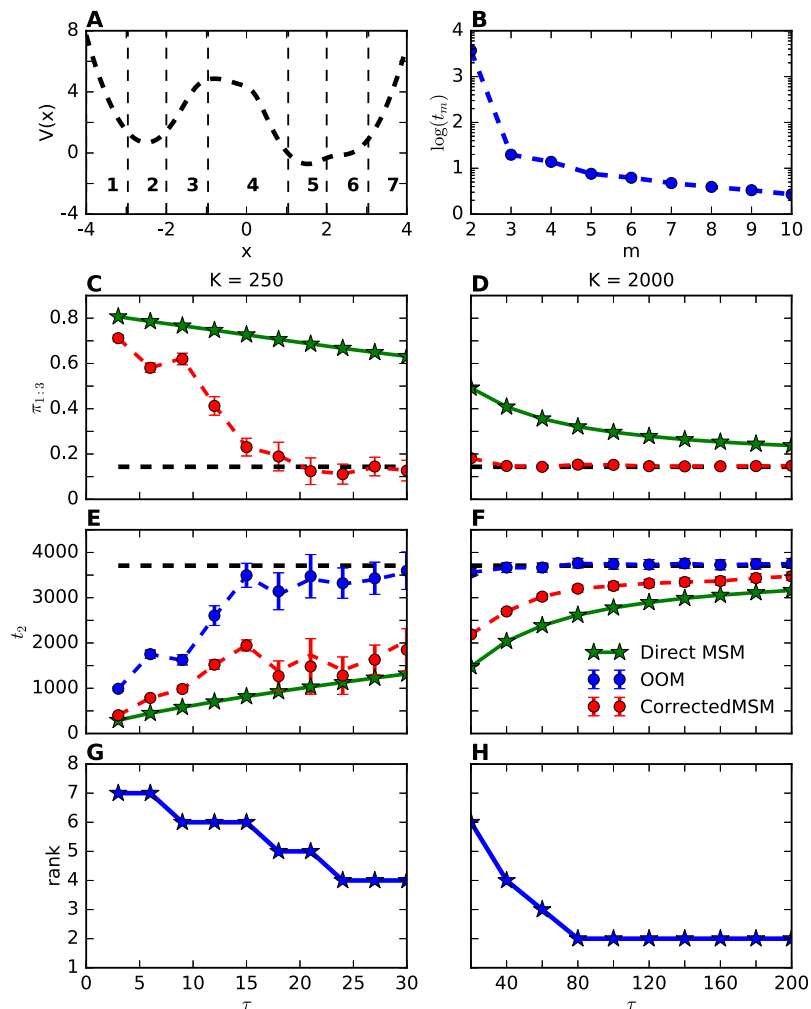
FIG. 5. (a) One-dimensional potential function and discretization of the landscape into seven states. (b) Decadic logarithm of the first nine implied time scales of the model system. ((c) and (d)) Estimates of the stationary probability of states 1-3 from the direct MSM (green) and the corrected MSM (red), compared to the reference (black dashed line). ((e) and (f)) Estimates of the slowest relaxation time scale t_2 from a direct MSM (green), the corrected MSM (red), and the OOM-based spectral estimation (blue), compared to the reference (black dashed line). ((g) and (h)) Model rank selected by the bootstrapping procedure. For all quantities derived from the OOM, the dashed lines indicate the estimated values using the complete data set, whereas the bullets and errorbars correspond to mean and standard error from the bootstrapping procedure. Note that the errorbars are hardly visible in panels (d) and (f).

uncertainties helps to select an appropriate model rank for each lag time, even when it is not obvious from the time scale plot. As expected, the system becomes effectively of rank 2 for lag times $\tau \geq 80$.

B. Molecular dynamics simulations of alanine dipeptide

Our second example is molecular dynamics simulation data of alanine dipeptide (Ac-A-NHMe) in explicit water. Alanine dipeptide has been used as a model system in numerous previous studies, see Refs. 25 and 35 and many others. It is well known that its dynamics can be described by the two-dimensional space of backbone dihedral angles ϕ, ψ. Fig. 6(a) shows the equilibrium probability distribution in this space with its three metastable minima in the upper left, central left, and central right parts of the plane. The slow dynamics consists of exchanges between the left and right parts ($t_2 \approx 1400$ ps) and between the two minima on the left ($t_3 \approx 70$ ps). We study the estimation of a Markov model using the discretization also indicated in panel (a) of Fig. 6. It was generated by k-means clustering of the data set described below using $N = 40$ cluster centers. We produced an ensemble of roughly 11 000 very short simulations of length 20 ps each. Simulations were initiated from eight different starting structures labelled by the numbers 1-8 in Fig. 6(b), see Appendix A for details.

It can be seen that the resulting empirical distribution does not even reach local equilibrium within the three metastable regions.

Like in the previous example, we find that it is possible to obtain precise estimates of stationary probabilities as soon as the convergence of the OOM-based time scales is achieved. In panel (c) of Fig. 6, we compare results for the equilibrium probability of all states in the right part of the plane, from a direct MSM and the corrected MSM. For lag times $\tau \geq 500$ fs, we are able to correct the bias introduced by strong non-equilibrium sampling.

In panels (d) and (f) of Fig. 6, we present estimates of the two slowest time scales t_2, t_3 produced by the same estimators as before (OOM in blue, direct MSM in green, and corrected MSM in red). Additionally, the cyan lines correspond to the time scale estimates of an MSM using equilibrium simulations and the same discretization (see Appendix A). We find that the OOM-based spectral estimation provides accurate time scale estimates for short lag times starting at $\tau = 500$ fs. Moreover, we notice that for lag times as small as these, MSM time scales are clearly lower than the true time scales, although a decent discretization is employed. The difference between OOM and MSM estimates indicates that an even finer discretization would be required to match the references at these lag times. The direct estimates, the reference equilibrium time
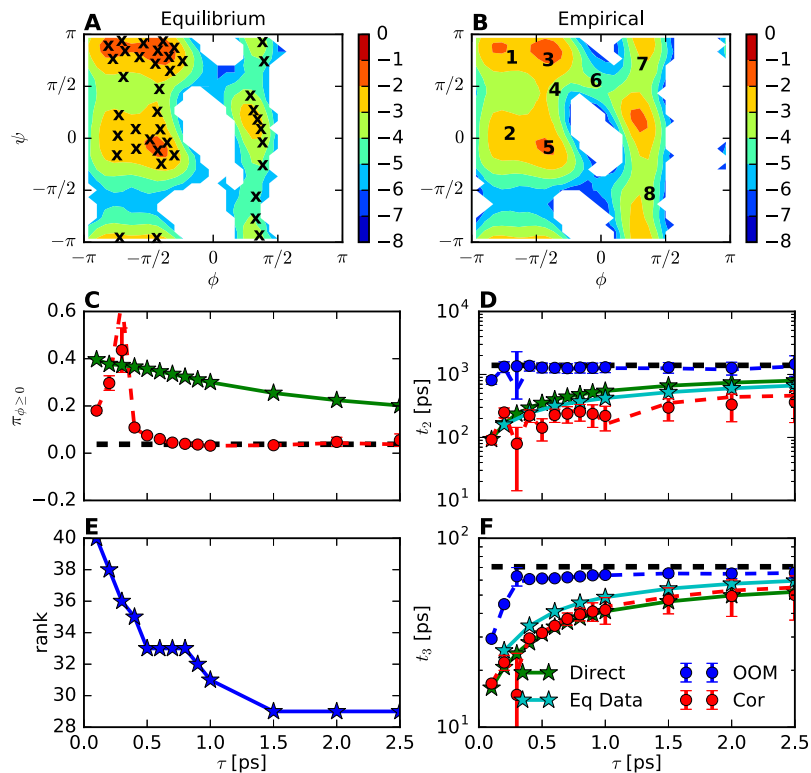
FIG. 6. Results for alanine dipeptide. (a) Equilibrium distribution (logarithmic scale) in the space of backbone dihedral angles ϕ, ψ and cluster centers of a forty-state k-means discretization used to analyze the data. (b) Empirical distribution (logarithmic scale) sampled by the data initiated from eight starting structures indicated by the numbers 1-8. (c) Equilibrium probability of all states in the right part of the plane estimated from the direct MSM (green) and the corrected MSM (red). Reference in black. (d) Estimates for the slowest relaxation time scale t_2 from a direct MSM (green), the corrected MSM (red), and the OOM-based estimation (blue). Reference values from equilibrium simulations are displayed in black. We also show the expected time scale estimate using the same forty-state discretization if equilibrium data were used (cyan line). (e) Model rank used for the OOM estimation as determined by the bootstrapping. (f) The same as (d) for the second slowest time scales t_3. For all quantities derived from the OOM, the dashed lines indicate the estimated values using the complete data set, whereas the bullets and errorbars correspond to mean and standard error from the bootstrapping procedure. Note that the errorbars are hardly visible in panels (c) and (f).

scales, and our OOM-based estimates of equilibrium time scales are nearly identical. Only the mean values extracted from bootstrapping for t_2 seem to be a bit low. This will be investigated further.

Finally, the selected model ranks shown in Fig. 6(e) confirm that our framework can work in situations where low-rank descriptions of the dynamics using only a few processes are not adequate.

C. Two-dimensional model system with poor discretization

Our final example is another finite state space Markov chain in the two-dimensional energy landscape shown in Fig. 7(a), defined by 40×40 microstates. Here we show the behavior of different estimators in an extreme case, where the discretization is so poor that MSM estimates fail completely. Transitions between neighboring states are now possible in both x- and y-directions, again based on a Metropolis criterion. We study the estimation of a Markov model using a discretization into 16 MSM states, also shown in Fig. 7(a). As can be seen in Fig. 7(b), there are two dominant time scales: $t_2 \approx 144\,000$ steps and $t_3 \approx 17\,000$ steps. The next time scale is clearly separated from the first two, after that, there is no more apparent time scale separation. This time, we fix the simulation length at $K = 5000$ steps, i.e., the trajectories are approximately 30 times shorter than the slowest time scale. The simulations are started from a uniform distribution over all microstates. In panels (c)-(h) of Fig. 7, we display the results if the number of simulations is set to $Q = 2000$ ((c), (e), and (g)) and $Q = 10\,000$ ((d), (f), and (h)).

In Figs. 7(c) and 7(d), we show the estimation results for the equilibrium probability of the states labeled 13, 14,

and 15. We expect it to be difficult to estimate this probability, as the states are blending different metastable regions and transition regions. It can be observed that the estimation of stationary probabilities is more sensitive to noise, see the results for $Q = 2000$. This observation is not surprising, as the stationary probabilities require accurate estimation of the two-step count matrices Eq. (7) from the data, which can be more difficult for rarely visited states. Still, for $Q = 10\,000$, a reliable estimate is achieved and the biased estimate of the direct MSM can be corrected. Another comparison we make is between the estimates from the corrected MSM and those from long equilibrium simulations that use the same number of total data points, i.e., $K = 2000 \cdot 5000 = 10^7$ for $Q = 2000$ and $K = 10\,000 \cdot 5000 = 5 \cdot 10^7$ for $Q = 10\,000$. We show mean values and standard errors from roughly 400 long simulations for $Q = 2000$, and roughly 900 simulations for $Q = 10\,000$. In both cases, the estimates from long equilibrium trajectories provide more accurate estimates. In practice, however, one needs to strike a balance between long trajectories that are more beneficial for the analysis and short trajectories that can be more efficient for sampling and state exploration.[36–38]

Again, we also compare the estimates for the slowest time scales t_2 ((e) and (f)) and t_3 ((g) and (h)) from a direct MSM, the corrected MSM, and the OOM-based spectral estimation. In both cases, correct estimates of both time scales can be obtained from the OOM, while both the direct and corrected MSMs estimate time scales one order of magnitude too small. This suggests that for a bad enough discretization, correcting for the effect of the non-equilibrium starting distribution will not be sufficient to achieve convergence in the time scales. However, the poor discretization quality is revealed by a large error between the OOM-based estimate and the
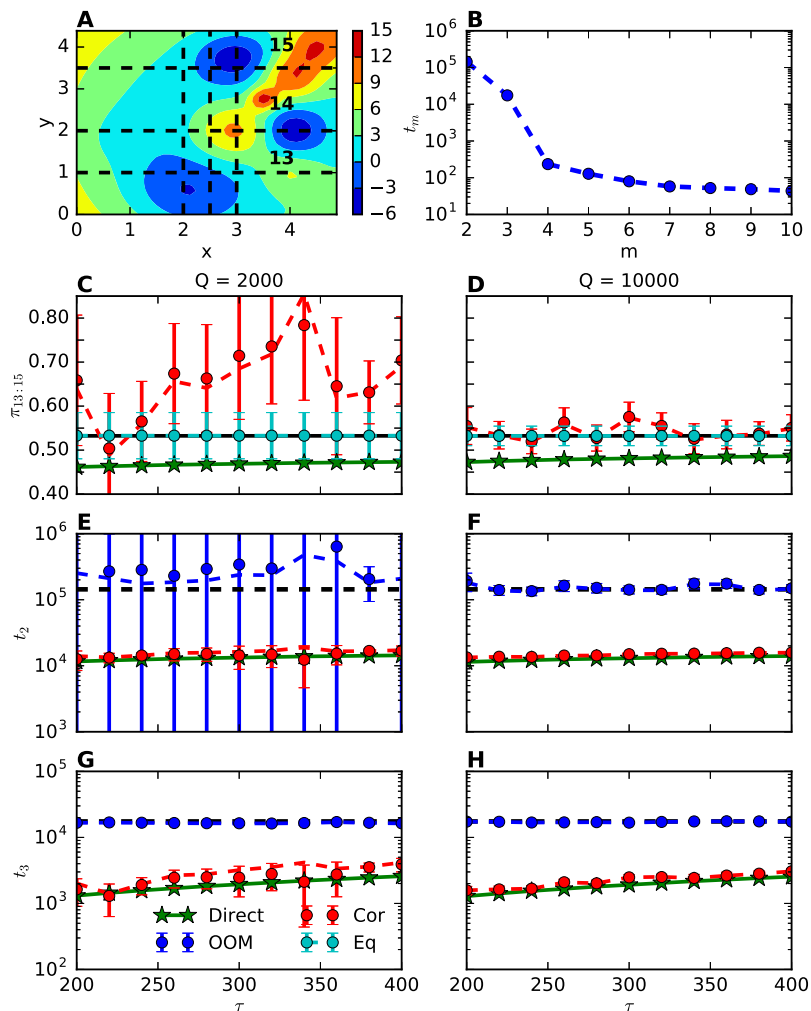
FIG. 7. (a) Two-dimensional potential function with discretization into 16 MSM states indicated by dashed lines. (b) Leading nine implied time scales t_m of the system. ((c) and (d)) Estimates of equilibrium probability of states 13, 14, and 15 from direct MSM (green) and the corrected MSM (red), compared to the reference (black line) and estimates from 400 (c) / 900 (d) different equilibrium simulations, shown by the cyan lines. ((e) and (f)) Estimates of slowest relaxation time scale t_2 from a direct MSM (green), the corrected MSM (red), and the OOM-based spectral estimation (blue), compared to the reference (black dashed line). ((g) and (h)) The same for t_3. For all quantities derived from the OOM, the dashed lines indicate the estimated values using the complete data set, whereas the bullets and errorbars correspond to mean and standard error from the bootstrapping procedure. Note that the errorbars are hardly visible in panels (f) and (h).

corrected MSM, and this observation can be exploited in order to improve the discretization and repeat the analysis.

V. CONCLUSIONS

We have investigated the quality of Markov state models when estimated from many simulations of short length, initiated from non-equilibrium starting conditions. We have derived an expression for the error between unbiased MSM transition probabilities and the expected estimate from many short simulations. This error is shown to depend on the simulation length, the lag time, and the state discretization. If ultra-long trajectories are employed, i.e., trajectories that are long compared to the slowest relaxation time scales, then the effect of the initial distribution is negligible and no further correction is needed. For ensembles of short trajectories, the situation is more complex. Preparing simulation trajectories in such a way that they emerge from a local equilibrium distribution does not appear to be of much practical use: this would only correct the first transition count of every trajectory while the subsequent trajectory segments are still biased. The local equilibrium will be lost for intermediate times along the trajectory as the trajectory ensemble is not in global equilibrium. In a similar sense discarding initial simulation fragments can reduce the bias but cannot systematically remove it. In particular, since the effect of the bias disappears with the slowest

relaxation times of the system, discarding pieces of simulation trajectories appears more harmful in terms of reducing the statistics than it is useful to reduce the bias. With the standard MSM estimator, the most effective and simplest method to reduce the bias from the initial trajectory distribution in fact seems to be using a longer lag time or a better state space discretization. These are already the usual objectives of MSM construction. However, if the discretization is poor, the estimation bias due to a non-equilibrium distribution can be dramatic at practically usable lag times.

The main result of this paper is that we propose an improved estimator of the MSM transition matrix which is not biased by the initial distribution. This new estimator is based on the theory of observable operator models. In contrast to the standard MSM estimator, the corrected MSM estimator does not only use the number of transitions observed between pairs of states at lag time τ but also the number of transitions at lag time 2τ. These statistics are combined to get a transition matrix estimate at lag time τ that is unbiased by the initial trajectory distribution. While it may seem that having to estimate statistics at 2τ is a deficiency compared to standard MSM estimation when only short simulation trajectories are available, please note that the corrected MSM estimator can get significantly better estimates at short lag times, so in practice the lag times needed for a converged MSM will be smaller than for the standard estimator.

Finally, we report a result from the OOM framework that shows how the model-free relaxation time scales can be computed from the same statistics used for the corrected MSM estimator (i.e., transition matrices at lag times τ and 2τ). These estimates are only impaired by statistical error but are not affected by systematic MSM error as no MSM is used in the process of obtaining them. The difference between the corrected MSM time scales and the OOM time scales can be used in order to assess the discretization quality, as this difference goes to zero in the limit of good discretization.

This paper addresses the long-standing controversy about the correct use of simulation data from short non-equilibrium simulations for MSM estimation and their effect on the estimation of equilibrium expectations and kinetics.

ACKNOWLEDGMENTS

APPENDIX A: SIMULATION SETUP OF ALANINE DIPEPTIDE

Molecular dynamics simulations of alanine dipeptide in explicit water at temperature 300 K were generated with ACEMD[39] software using the AMBER ff-99SB-ILDN force field[40] and an integration time step of 2 fs. The peptide was simulated inside a cubic box of volume $(2.3222 \text{ nm})^3$ containing 651 TIP3P water molecules. The Langevin thermostat was used. The electrostatics were computed every two time steps by the particle-mesh Ewald (PME) method,[41] using real-space cutoff 0.9 nm and grid spacing 0.1 nm. All bonds between hydrogens and heavy atoms were constrained.

We have produced 11 388 ultra short simulations of length 20 ps each, with 50 fs saving interval. The simulations were initiated from eight different structures; their projections into $\phi - \psi$-space are indicated by the numbers 1-8 in Fig. 6(b). The probabilities to start in each of these structures are given by the vector

$$\rho_1 = \begin{bmatrix} 0.05 & 0.05 & 0.2 & 0.2 & 0.2 & 0.1 & 0.1 & 0.1 \end{bmatrix}. \quad (A1)$$

These simulations were used to perform the analyses described in Sec. IV B. Using the same setup, we produced 2363 long runs of 1 ns simulation time each, with 1 ps saving interval. We estimated a Markov model on the 40-state k-means discretization at lag time $\tau = 100$ ps using this data set, and extracted the reference time scales and equilibrium probabilities shown as black lines in Fig. 6. Also, we used the stationary probabilities estimated from this model to initialize 203 short equilibrium runs of 500 ps simulation time each, with 100 fs saving interval. This data set was used to compute the equilibrium time scales of the k-means discretization shown as cyan lines in Figs. 6(d) and 6(f).

APPENDIX B: OOM PROBABILITY OF OBSERVATION SEQUENCE

Here, we show the derivation of the path probability formula Eq. (36), that can also be found in Ref. 21. In general, the left-hand side of Eq. (36) can be expressed by repeated integrals over the transition kernel,

$$\mathbb{P}(X_\tau \in A_1, \ldots, X_{l\tau} \in A_l) = \int_\Omega \int_{A_1} \ldots \int_{A_l} \mathrm{dx}_0 \ldots \mathrm{dx}_l \, \pi(x_0) p(x_0, x_1; \tau) \ldots p(x_{l-1}, x_l; \tau). \quad (B1)$$

Note that π appears in the first integral as we assumed that the dynamics is in equilibrium, i.e., the initial distribution equals π. Next, we replace all transition kernels by the expansion in Eq. (4),

$$\mathbb{P}(X_\tau \in A_1, \ldots, X_{l\tau} \in A_l) = \sum_{m_0=1}^{M} \sum_{m_1=1}^{M} \cdots \sum_{m_{l-1}=1}^{M} \left[\int_\Omega \mathrm{dx}_0 \, \pi(x_0) \psi_{m_0}(x_0) \right] \lambda_{m_0}(\tau)$$

$$\times \left[\int_{A_1} \mathrm{dx}_1 \, \psi_{m_0}(x_1) \pi(x_1) \psi_{m_1}(x_1) \right] \ldots \lambda_{m_{l-1}}(\tau) \left[\int_{A_l} \mathrm{dx}_l \, \psi_{m_{l-1}}(x_l) \pi(x_l) \right] \quad (B2)$$

$$= \sum_{m_0=1}^{M} \sum_{m_1=1}^{M} \cdots \sum_{m_{l-1}=1}^{M} \delta_{1,m_0} \Xi_{A_1}(m_0, m_1) \ldots \Xi_{A_l}(m_{l-1}, 1). \quad (B3)$$

In the second equation, we have used the π-orthogonality of the eigenfunctions ψ_{m_0} and the fact that $\psi_1 \equiv 1$ in order to replace the x_0-integral by δ_{1,m_0}. For the last integral, we have also used that $\psi_1 \equiv 1$. This is a sequence of matrix-vector products. It remains to use $\delta_{1,m_0} = \omega(m_0)$ and that $\Xi_{A_l}(m_{l-1}, 1) = [\Xi_{A_l}\sigma](m_{l-1})$. In matrix notation, Eq. (36) follows:

$$\mathbb{P}(X_\tau \in A_1, \ldots, X_{l\tau} \in A_l) = \omega^T \Xi_{A_1} \ldots \Xi_{A_l} \sigma \quad (B4)$$

Finally, note that this derivation also works if the dynamics is not in equilibrium. In this case, the vector ω is given by $\omega(m_0) = \int_\Omega \mathrm{dx}_0 \, \rho_0(x_0) \psi_{m_0}(x_0)$, where ρ_0 is the non-equilibrium initial condition.

APPENDIX C: VARIABLE SIMULATION LENGTH

Here, we verify that the estimation algorithm from Sec. III B can be applied to data sets comprising simulations of

non-uniform length. We assume that for $j = 1, \ldots, J$, there is an ensemble of Q_j simulations of length $K_j + 2\tau$, i.e., K_j transition pairs/triples will be used from each of these trajectories. We assume that $Q_j \to \infty$ for all j, such that every sub-ensemble samples from an empirical distribution ρ_j. Define the number of data points generated by the jth ensemble as $T_j = Q_j K_j$ and the total number of data points by

$$T := \sum_{j=1}^{J} Q_j K_j. \tag{C1}$$

Moreover, we assume that $\frac{T_j}{T} \to \alpha_j$, i.e., the fraction of data points generated by the jth ensemble approaches a constant for all j. Let us define the distribution

$$\rho = \sum_{j=1}^{J} \alpha_j \rho_j. \tag{C2}$$

Trajectories of length $K_j + 2\tau$ are enumerated by q_j and labelled $\mathbf{Y}_{q_j}$. Further, let $s_{K_j}(\mathbf{Y}_{q_j})$ be any of the estimators from Eqs. (5)–(7), where the subscript K_j indicates that $K - 2\tau$ in Eqs. (5)–(7) must be replaced by K_j. In addition, denote by $s(\mathbf{Y}_{q_j})$ the same estimator but without the normalization. Also, let c_{ρ_j} denote the corresponding correlation from Eqs. (20), (21), and (43) with respect to the density ρ_j. It follows that

$$\bar{s}_T := \frac{1}{T} \left[\sum_{q_1=1}^{Q_1} s(\mathbf{Y}_{q_1}) + \cdots + \sum_{q_J=1}^{Q_J} s(\mathbf{Y}_{q_J}) \right] \tag{C3}$$

$$= \frac{T_1}{T} \left[\frac{1}{T_1} \sum_{q_1=1}^{Q_1} s(\mathbf{Y}_{q_1}) \right] + \cdots + \frac{T_J}{T} \left[\frac{1}{T_J} \sum_{q_J=1}^{Q_J} s(\mathbf{Y}_{q_J}) \right] \tag{C4}$$

$$= \frac{T_1}{T} \left[\frac{1}{Q_1} \sum_{q_1=1}^{Q_1} s_{K_1}(\mathbf{Y}_{q_1}) \right] + \cdots + \frac{T_J}{T} \left[\frac{1}{Q_J} \sum_{q_J=1}^{Q_J} s_{K_J}(\mathbf{Y}_{q_J}) \right] \tag{C5}$$

$$\to \alpha_1 \mathbb{E}\left(s_{K_1}\right) + \cdots + \alpha_J \mathbb{E}\left(s_{K_J}\right) \tag{C6}$$

$$= \alpha_1 c_{\rho_1} + \cdots + \alpha_J c_{\rho_J} \tag{C7}$$

$$= c_\rho. \tag{C8}$$

The convergence in Eq. (C6) is convergence in probability. Thus, if we sum up all visits/transitions/two-step transitions, and divide by the total number of data points in the end, we arrive at an asymptotically correct estimator of the correlations with respect to the density ρ. As the OOM estimation algorithm only relies on consistent estimators for correlations with respect to some empirical density ρ, it can still be applied in this setting. Finally, the normalization by $\frac{1}{T}$ can be omitted in practice because it cancels out in Eqs. (47)-(52).

[1] 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).

[2] C. Schütte and M. Sarich, *Metastability and Markov State Models in Molecular Dynamics, Modeling, Analysis, Algorithmic Approaches*, Courant Lecture Notes Vol. 24 (American Mathematical Society and Courant Institute of Mathematical Sciences, 2013).

[3] *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, edited by G. R. Bowman, V. S. Pande, and F. Noé (Springer, Netherlands, 2014).

[4] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid Monte Carlo," Comput. Phys. **151**, 146–168 (1999).

[5] W. C. Swope, J. W. Pitera, and F. Suits, "Describing protein folding kinetics by molecular dynamics simulations. 1. Theory," J. Phys. Chem. B **108**(21), 6571–6581 (2004).

[6] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, "Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states," J. Chem. Phys. **126**, 155102 (2007).

[7] 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).

[8] F. Noé, "Probability distributions of molecular observables computed from Markov models," J. Chem. Phys. **128**(24), 244103 (2008).

[9] N. V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," J. Phys. Chem. B **112**, 6057–6069 (2008).

[10] 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**, 124101 (2009).

[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. **106**, 19011–19016 (2009).

[12] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9," J. Am. Chem. Soc. **132**(5), 1526–1528 (2010).

[13] W. Zhuang, R. Z. Cui, D.-A. Silva, and X. Huang, "Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the Markov state model approach," J. Phys. Chem. B **115**, 5415–5424 (2011).

[14] M. Sarich, F. Noé, and C. Schütte, "On the approximation quality of Markov state models," Multiscale Model. Simul. **8**, 1154–1177 (2010).

[15] J. Chodera, V. S. Pande, M. Weber, F. Noé, and C. Schütte, personal communication.

[16] S. Röblitz, "Statistical error estimation and grid-free hierarchical refinement in conformation dynamics," Ph.D. thesis, Freie Universität Berlin, 2009.

[17] M. Weber, "A subspace approach to molecular Markov state models via an infinitesimal generator," Technical Report 09–27, ZIB, 2009.

[18] X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande, "Rapid equilibrium sampling initiated from nonequilibrium data," Proc. Natl. Acad. Sci. **106**(47), 19765–19769 (2009).

[19] G.R. Bowman, X. Huang, and V. S. Pande, "Using generalized ensemble simulations and Markov state models to identify conformational states," Methods **49**(2), 197–201 (2009).

[20] H. Jaeger, "Observable operator models for discrete stochastic time series," Neural Comput. **12**(6), 1371–1398 (2000).

[21] H. Wu, J.-H. Prinz, and F. Noé, "Projected metastable Markov processes and their estimation with observable operator models," J. Chem. Phys. **143**(14), 144101 (2015).

[22] H. Wu and F. Noé, "Spectral learning of dynamic systems from nonequilibrium data," in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 4179–4187.

[23] N. Djurdjevac, M. Sarich, and C. Schütte, "Estimating the eigenvalue error of Markov state models," Multiscale Model. Simul. **10**(1), 61–81 (2012).

[24] F. Noé and F. Nüske, "A variational approach to modeling slow processes in stochastic dynamical systems," Multiscale Model. Simul. **11**, 635–655 (2013).

[25] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, "Variational approach to molecular kinetics," J. Chem. Theory Comput. **10**, 1739–1752 (2014).

[26] M. Weber, "Meshless methods in conformation dynamics," Ph.D. thesis, Freie Universität Berlin, 2006.

[27] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, "Estimation and uncertainty of reversible Markov models," J. Chem. Phys. **143**(17), 174101 (2015).

[28] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, "Atomic-level characterization of the structural dynamics of proteins," Science **330**, 341–346 (2010).

[29] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," Science **334**(6055), 517–520 (2011).

[30] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, "Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules," J. Chem. Phys. **139**, 184114 (2013).

[31] J.-H. Prinz, "Advanced estimation methods for Markov models of dynamical systems," Ph.D. thesis, Freie Universität Berlin, 2012.

[32] G. W. Stewart, "Perturbation theory for the singular value decomposition," in *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, edited by R. J. Vaccaro (Elsevier, 1991).

[33] F. Noé, "Statistical inefficiency of Markov model count matrices" (unpublished).

[34] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A software package for estimation, validation, and analysis of Markov models," J. Chem. Theory Comput. **11**, 5525 (2015).

[35] F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller, "Dynamic properties of force fields," J. Chem. Phys. **142**(8), 084101 (2015).

[36] J. Preto and C. Clementi, "Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics," Phys. Chem. Chem. Phys. **16**, 19181–19191 (2014).

[37] S. Doerr and G. De Fabritiis, "On-the-fly learning and sampling of lig- and binding by high-throughput molecular simulations," J. Chem. Theory Comput. **10**(5), 2064–2069 (2014).

[38] S. Doerr, M. J. Harvey, F. Noé, and G. De Fabritiis, "HTMD: High-throughput molecular dynamics for molecular discovery," J. Chem. Theory Comput. **12**(4), 1845–1852 (2016).

[39] M. J. Harvey, G. Giupponi, and G. De Fabritiis, "ACEMD: Accelerating biomolecular dynamics in the microsecond time scale," J. Chem. Theory Comput. **5**(6), 1632–1639 (2009).

[40] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," Proteins: Struct., Funct., Bioinf. **78**(8), 1950–1958 (2010).

[41] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N -log(N) method for Ewald sums in large systems," J. Chem. Phys. **98**, 10089–10092 (1993).