REVERSIBLE MARKOV STATE MODELS

BENJAMIN TRENDELKAMP-SCHROER

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften

Fachbereich Mathematik und Informatik
Freie Universität Berlin

Juni 2016

ABSTRACT

The dynamics of proteins is characterized by infrequent complex rearrangements of the overall protein structure, so called conformational changes. Experiments cannot observe conformational changes directly with atomistic resolution so that computer simulations have to be used for the detailed study of these processes. The rare occurrence of conformational changes on timescales accessible by simulations results in a sampling problem. Obtaining reliable estimates for quantities connected to rare transition events from simulation data is very difficult and can require prohibitively long running times for computations. Markov state models can systematically integrate ensembles of short trajectories and thus effectively parallelize the computational effort, but the rare transition events still need to be spontaneously sampled in the data.

Reversible Markov state models can be used to overcome the sampling problem. Reliable estimates of probabilities for rare transitions can be computed with a simulation effort that is orders of magnitudes smaller than the average effort required to observe even a single rare event. They arise naturally as a discrete space, finite data description of metastable conformational processes.

In this thesis methods for the estimation and uncertainty quantification of reversible Markov state models and a novel approach for the estimation of transition probabilities from simulation data containing rare events are developed. Estimators for reversible transition probabilities are constructed using a maximum likelihood approach and efficient algorithms for the solution of the arising optimization problem are developed. The method can also be applied in situations in which additional information about stationary probabilities is available. Uncertainty quantification for reversible Markov state models is facilitated using a Bayesian approach. An efficient Monte Carlo algorithm for sampling of reversible transition matrices from the arising posterior distribution is developed. A variant that can incorporate *a priori* information about the stationary probabilities is also presented. Stationary probabilities can often be efficiently estimated from enhanced sampling simulations which do not suffer from a sampling problem. For many interesting metastable processes only one direction of the process is rare while the reverse direction can be sampled efficiently for appropriately chosen initial conditions. Reversibility can be used to estimate transition probabilities in situations in which only one direction of the process of interest has been observed if information about the stationary probabilities of the process is available. So that enhanced sampling simulations and short molecular dynamics trajectories can be systematically combined using reversible Markov state models.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

[1] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, 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." In: *J. Chem. Theory Comput.* (2015). URL: http://dx.doi.org/10.1021/acs.jctc.5b00743.

[2] M. Senne, B. Trendelkamp-Schroer, A. S. Mey, C. Schütte, and F. Noé. "EMMA: A software package for Markov model building and analysis." In: *J. Chem. Theory Comput.* 8.7 (2012), pp. 2223–2238. URL: http://dx.doi.org/10.1021/ct300274u.

[3] B. Trendelkamp-Schroer and F. Noé. "Efficient Bayesian estimation of Markov model transition matrices with given stationary distribution." In: *J. Chem. Phys.* 138.16 (2013), p. 164113. URL: http://scitation.aip.org/content/aip/journal/jcp/138/16/10.1063/1.4801325.

[4] B. Trendelkamp-Schroer and F. Noé. "Efficient Estimation of Rare-Event Kinetics." In: *Phys. Rev. X* 6 (1 2016), p. 011009. URL: http://link.aps.org/doi/10.1103/PhysRevX.6.011009.

[5] B. Trendelkamp-Schroer, H. Wu, and F. Noé. *Reversible Markov chain estimation using convex-concave programming.* 2016. URL: http://arxiv.org/abs/1603.01640.

[6] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. "Estimation and uncertainty of reversible Markov models." In: *J. Chem. Phys.* 143.17 (2015), p. 174101. URL: http://scitation.aip.org/content/aip/journal/jcp/143/17/10.1063/1.4934536.

Fortuna imperatrix mundi

— *C. Orff*

ACKNOWLEDGMENTS

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

CLT   central limit theorem

DHAM   dynamical histogram reweighting analysis method

dTRAM   discrete transition matrix reweighting analysis method

DNA   deoxyribonucleic acid

GPU   graphics processing unit

HIV   human immunodeficiency virus

KKT   Karush-Kuhn-Tucker

MCMC   Markov chain Monte Carlo

MD   Molecular dynamics

MFPT   mean first-passage time

MHC   major histocompatibility complex

MINRES   minimum residual

MLE   maximum likelihood estimator

MSM   Markov state model

NMR   nuclear magnetic resonance

RHS   right-hand side

RNA   ribobucleic acid

SDE   stochastic differential equation

TPT   transition path theory

VI   variational inequality

WHAM   weighted histogram analysis method

INTRODUCTION

Life emerges from a complex interplay of different biological processes taking place on multiple scales. On the smallest scales all necessary tasks are carried out by proteins - microscopic molecular machines which perform a diverse set of functions. They aid in the production and degradation of substances, are involved in transportation and signal transduction, generate mechanical forces from chemical reactions, or clog together into stiff fibers to stabilize cells.

Proteins consist of thousands of atoms. They are sequences of amino acids that fold into a specific three-dimensional structure - the protein's native state. The native state is not static. Long lived substructures, so called conformations, are present in the native state. Random fluctuations between conformations occur due to thermal activation. Conformations correspond to regions of low energy which are separated by energetic barriers. The waiting time in one of the conformation increases exponentially with the barrier height so that conformational changes are rare events, cf. Figure 1.1. Conformational changes are important for a protein's function and its interaction with other molecules.

Conformational changes are complex processes involving large scale rearrangements of the overall protein geometry and the forming and breaking apart of substructures such as helices and beta sheets. Experiments cannot probe proteins simultaneously in time *and* space on scales necessary to resolve the interesting processes. Computer simulations can achieve such a resolution, but they suffer from a severe sampling problem. The elementary timestep for the integration of the equations of motions is on the order of femtoseconds ($10^{-15}s$) while transitions between conformations occur on the order of microseconds ($10^{-6}s$) or milliseconds ($10^{-3}s$). Trajectories that sufficiently sample the rare switching between conformations cannot be computed even by the most powerful computers. Conformational changes often involve multiple transition pathways with many long-lived intermediates so that they cannot be described in terms of a simple reaction with a single transition state.

Markov state models (MSM) describe conformation dynamics by a directed network in which vertices (states) correspond to conformations and edges correspond to conformational transitions. The probability for a transition is encoded into the weight of the corresponding edge. The network is modeled and parameterized using simulation data. Parameters for the network can be estimated using many short trajectories which can be simulated in parallel. Still, reliable values for the transition probabilities can only be obtained from the trajectories, if all steps
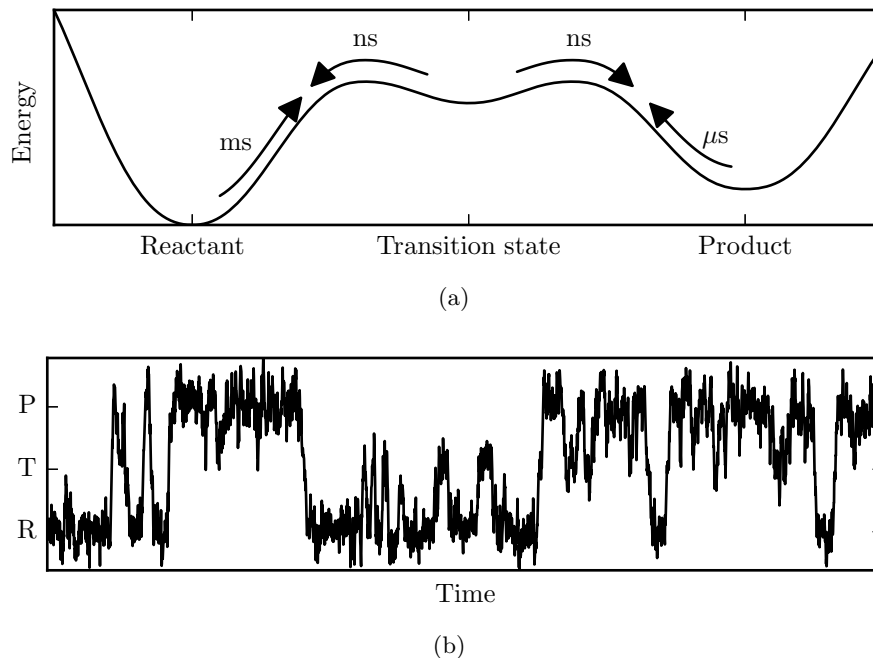
Figure 1.1: Conformation dynamics a) Conformations correspond to regions of low energy. Arrows indicate the average waiting time for a transition. b) Conformational trajectory, random fluctuations are due to thermal activation, transitions between conformations are rare events.

of the forward *and* the reverse process have been sampled, so that the network is connected in the forward and in the reverse direction.

Invariance of the equations of motions with respect to time-reversal results in a detailed balance equation for the probabilities of the forward and the reverse transition. This detailed balance condition is *not* satisfied if transition probabilities are estimated from relative frequencies of transitions observed in the data. But, estimators that obey the detailed balance equation lead to optimization problems which are difficult to solve.

It is shown that estimators which obey the detailed balance equation can be constructed if the transition probabilities are interpreted as parameters of a statistical model. The parameters can be computed using a maximum likelihood approach. An efficient algorithm for the solution of the arising optimization problem is developed. The proposed algorithm is superlinearly convergent, robust with respect to the input data, and can be applied in situations in which the MSM contains a large number of states.

Any data-driven approach needs to quantify the uncertainty of estimated values. The parameters for the MSM are often estimated from data that contains rare events so that estimates for their uncertainty cannot be obtained using standard methods which are only valid asymptotically. Other inference methods which require to sample the set of transition probabilities according to a high-dimensional probability dis-

tribution have to be used. But, this is difficult for transition probabilities which have to fulfill a detailed balance condition.

Uncertainty quantification for reversible MSMs is possible if a Bayesian viewpoint is adopted. A probability distribution (posterior distribution) for reversible transition probabilities is constructed and a Monte Carlo algorithm is developed to sample reversible transition matrices from the posterior. The algorithm is efficient even in situations in which the MSM contains a large number of states and computes reliable estimates for data containing rare events.

The detailed balance condition can be used to efficiently estimate transition probabilities in situations in which only one direction of the process has been observed *and* stationary probabilities for the states of the MSM are available. Stationary probabilities can be efficiently estimated from enhanced sampling simulations which do not suffer from a sampling problem. But, enhanced sampling simulations cannot be used to estimate transition probabilities. A systematic approach for the estimation of probabilities for rare transitions is outlined. Reversible MSMs can be used to combine enhanced sampling simulations and short relaxation trajectories generated by standard molecular dynamics (MD) simulations. Reliable estimates of transition probabilities can be obtained orders of magnitude before a single rare event would have been observed on average.

Three examples for proteins adapted from [6, 41] are shown in Figure 1.2 and briefly discussed. The beta-2 adrenergic receptor (Figure 1.2a [40]) is located on the cell membrane and reacts to adrenaline molecules released into the blood. When the receptor is stimulated by an adrenaline molecule it undergoes a conformational change signaling the cell to increase its activity leading, for example, to an increased heart-rate. A beta blocker (shown in yellow) occupying the adrenalin receptor site can be used to treat cardiac arrhythmia and can help to protect from heart attacks.

The major histocompatibility complex (MHC-I, Figure 1.2b [39]) functions as a cell's fingerprint. The MHC-I presents short fragments (in red) of all proteins, which have been broken apart inside the cell, on the cell-surface. The immune system can then sense infections and abnormal cell-growth by scanning the fragments presented on the surface. The MHC-I needs to bind and release these fragments with the right affinity, so that the immune system can successfully read out the cell's fingerprint.

The human immunodeficiency virus (HIV) reverse transcriptase molecule (Figure 1.2c [38]) is used by HIV to insert genetic information into the genome of the host cell. The genetic information of the virus is encoded into ribonucleic acid (RNA), but the host cell's genetic information is encoded into deoxyribonucleic acid (DNA). The reverse transcriptase molecule can assemble a DNA counterpart to a piece of viral RNA, which can then be inserted into the host cell's genome. As-

(a) Adrenergic receptor   (b) MHC-I   (c) HIV-transcriptase
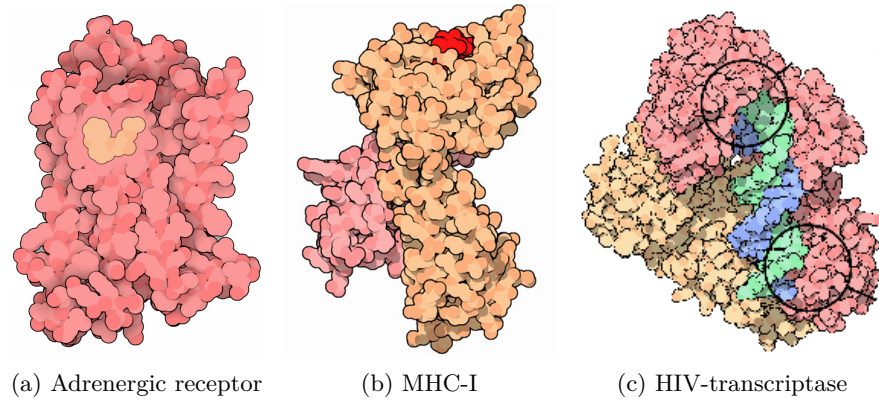
Figure 1.2: Examples for proteins: HIV, MHC-I, and adrenalin-receptor.

sembly of viral DNA (green and blue) is performed in the claw-shaped active site. Occupying or inducing conformational distortions in the active site can bring the process of reverse transcription to a halt, slowing the spread of the virus [125].

Evidence for the existence of conformations was found in experiments probing the rebinding of carbon monoxide and dioxygen to myoglobin after their removal by photodissociation. A conformation dependent rebinding rate was able to explain the measured nonexponential decay of unbound myoglobin indicating the existence of different conformational states for bound myoglobin molecules in the ensemble prior to the photodissociation pulse [3]. X-ray diffraction measurements and Mössbauer experiments found evidence for conformational flexibility in proteins in the crystal state or in frozen solution [3, 36, 45, 53]. Nuclear magnetic resonance (NMR) measurements revealed conformational flexibility on the millisecond timescale [52].

A computational study of myoglobin [29] demonstrated that the native state of a protein consists of many local minima separated by barriers. The existence of a sampling problem in molecular dynamics simulations was pointed out in [19]. Many different approaches have been developed to alleviate the sampling problem. A rigorous mathematical approach which allows to address the sampling problem for conformation dynamics is the transfer operator approach [94, 95].

The atomistic motion of a protein subject to thermal fluctuations can be described by a Markov process. A Markov process is a stochastic process which instantly forgets its past. For molecular dynamics this process is metastable, which means that the process tends to stay in a subset of the state space for a long time before transitioning to another subset in which it will again remain for a long time. The average time for a transition is much shorter than the average time the process spends in one of the metastable sets. Protein conformations correspond to metastable sets of the dynamics.
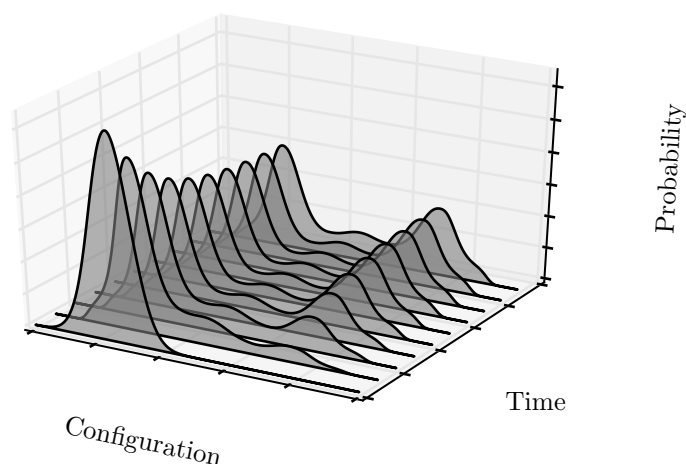
Figure 1.3: Probability transport by conformations dynamics. Part of the probability that is initially localized in the left well (reactant) is shifted to the right well (product). At large times the density becomes invariant under the dynamics (convergence to the equilibrium density).

As a result of metastability a finite length trajectory will contain very few or no transitions between metastable sets. It will be impossible to compute reliable values for the equilibrium probabilities from such trajectories. The total time that the trajectory spends in each conformation will depend sensitively on the starting point so that the relative duration of stays cannot serve as a good approximation for the equilibrium probabilities. Probabilities for transitions between conformations cannot be estimated reliably since the trajectory contains only very few or no transition events. Such a trajectory is called non-ergodic since averages over the state space cannot be approximated by time-averages obtained from the trajectory.

As ergodic trajectories are computationally unavailable for metastable systems a different approach to the computation of conformation dynamics has to be used. Single trajectories are replaced by statistical ensembles which can be described by probability densities. Evolution of the ensemble results in a transport of probability, cf. Figure 1.3. A probability density that is supported on a metastable set will remain almost invariant under the action of the dynamics on timescales smaller than the characteristic timescale for the conformational process. For long times any initial probability distribution is transported to the equilibrium distribution which is invariant under the action of the dynamics.

Transport of probability densities can be described by a transfer operator Figure 1.4. The transfer operator is a bounded linear mapping on an infinite dimensional space of suitable functions. Conformations are sets whose characteristic functions are almost invariant under the action of the transfer operator. Eigenvectors of the transfer operator
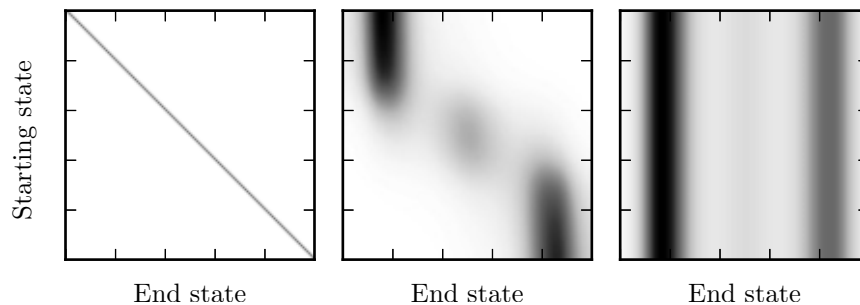
Figure 1.4: Transfer operator. Matrix representation of the operator, large values are dark, small values are light. Each row is the probability density for finding the system in an end state after a fixed lagtime, given that the system was started in the state corresponding to the row index. At short times (left) the probability for transitioning to an end state is approximately normally distributed around the starting state. At longer times (middle) metastability of the dynamics is reflected by a block structure of the operator. Blocks correspond to conformations. For very long times (right) each row is given by the equilibrium distribution of the dynamics.

corresponding to its dominant eigenvalues are almost constant on sets corresponding to molecular conformations and they can be used to identify them [23, 24].

Unfortunately the transfer operator for molecular dynamics is computationally inaccessible in full detail. But, the operator can be discretized and the resulting matrix representation can be used to identify molecular conformations and to extract interesting properties of the dynamics. Matrix elements of the discretized operator can be interpreted as transition probabilities between sets if the process is reversible and the space of molecular configurations is discretized by a finite number of nonoverlapping sets covering the space. Such a discretization of the transfer operator is called a Markov state model, if the memory that is introduced by the discretization of the process is removed by choosing an appropriate lagtime. The matrix elements for the MSMs can be estimated from an ensemble of short trajectories whose initial configurations are distributed according to the stationary probabilities of the dynamics [95]. Ensembles of short trajectories can be efficiently simulated in parallel.

If the dynamics is reversible then the transfer operator is as a self-adjoint operator on a suitably weighted space of square integrable functions [95]. Suitable weights are the equilibrium probabilities for the dynamics. For molecular dynamics they are given by the canonical distribution, i.e. the probabilities for molecular configurations in thermal equilibrium with a heat bath.

Other approaches in which probabilities for conformational transitions are estimated using an ensemble of short trajectories were also developed. In [72, 75, 76] conformation dynamics was modeled by a weighted network with edges representing the transition between sta-

ble states of the system. The complexity of large scale conformational changes could then be addressed by computing the best paths on the network going from the reactant to the product state.

Transition path theory (TPT) [119] provides a rigorous framework to study the statistical properties of reactive trajectories, which can be used to study conformation dynamics. In [77] it was shown that a combination of the transfer operator approach and transition path theory using Markov state models made it possible to compute the interesting folding pathways of a protein using only short simulations with starting points that were not distributed according to the equilibrium weights of the system.

The possibility to attack conformation dynamics by a massively distributed approach led to the development of the Folding@home project [98, 100, 103] which allows users to donate idle processing time on their home computers to scientists performing simulations of proteins. Markov state models have been established as the unifying theoretical framework for all approaches in which ensembles of molecular simulations were used for computations of conformation dynamics [70, 82, 84, 93, 94, 102, 105], see [11] for a comprehensive survey.

Transition path sampling [10, 21], transition interface sampling [30], and milestoning [33] can also be used to compute transition paths from short simulation trajectories. In contrast to global approaches like MSMs who construct a network of conformational transitions from an ensemble of trajectories these methods are local. They need *a priori* information about the transition end-states, and the transition path needs to be resolved by the reaction coordinate, cores, or milestones, which have to be defined beforehand.

Another approach to rare event sampling based on optimal control of diffusion processes was outlined in [44]. An optimal nonequilibrium forcing for the diffusion process is computed starting from an initial guess. The optimal forcing can be used to estimate rare event probabilities with zero statistical error. The method requires the *a priori* definition of starting states and target sets and the nonequilibrium forcing has to be approximated by a finite number of ansatz functions. In [120] the optimal nonequilibrium forcing is computed via local solutions of a partial differential equation associated to the optimal control problem.

Enhanced sampling methods can achieve an enormous speed up when computing equilibrium expectations. They simulate the system using an artificial dynamics that equilibrates more rapidly than the original one. In umbrella sampling [110] the system is 'dragged' over energy barriers along a reaction coordinate parameterizing the interesting transitions. Replica exchange dynamics [104] couple the unbiased ensemble with ensembles at higher temperatures at which the rare events occur more frequently. Metadynamics [42, 59] fill up the free energy landscape according to the frequency of visits by the evolving trajectory. The unbiased equilibrium probabilities can then be computed from the

biased probabilities using statistical reweighting methods [5, 35, 57, 99, 106]. The dynamical properties of the unbiased ensemble, such as transition rates, relaxation time scales, and transition pathways however cannot be estimated from the biased simulations.

The dynamical histogram reweighting analysis method (DHAM) [90] and the discrete transition matrix reweighting analysis method (dTRAM) [123, 124] enable the combination of enhanced sampling simulations and standard (unbiased) molecular dynamics simulations into an estimator for the transition probabilities of the unbiased ensemble which can the be used to compute interesting dynamical properties like transition rates, relaxation time scales, and transition pathways.

Computing statistical errors for MSMs is important since MSMs are usually estimated from trajectories containing rare events. In [77, 101] error estimates were obtained using the posterior ensemble of nonreversible transition matrices. This approach was extended to the posterior ensemble of reversible transition matrices [4, 63, 65, 73]. Perturbation approaches for the nonreversible posterior are outlined in [46, 88].

The probabilistic interpretation of the transfer operator in [95] requires that the underlying Markov process is reversible. This property is then carried over to the transition matrix of the Markov state model. But the standard approach of estimating this matrix from simulation data results in a transition matrix that is *not* reversible. Existing methods that use the transition matrix for a spectral clustering of the state space [89] and the computation of dominant transition pathways [64] do not require reversible transition matrices. But, a proof that spectral clusters correspond to almost invariant sets has only been given in the reversible case [48]. A reversible estimator can also have better asymptotic variance than the standard estimator [85]. The computation of eigenvalues and eigenvectors of reversible matrices is also more robust than the computation of eigenvalues and eigenvectors for nonreversible matrices.

Maximum likelihood estimators that ensure reversibility lead to optimization problems that are solved by a self-consistent iteration [12, 84, 123]. The iteration can take a long time to converge and the convergence behavior can be sensitive to the data that is used.

This thesis discusses efficient algorithms to estimate reversible transition matrices given observations of a Markov process. It is shown that a duality argument from [123] can be used to transform the maximum likelihood estimation problem into a constrained saddle-point problem. The resulting saddle-point problem is then solved using a primal-dual interior-point algorithm which can efficiently compute the maximum likelihood estimator (MLE) for Markov chains with a large number of states and is more robust than the previously used self-consistent iteration. The algorithmic approach allows to tackle a number of problems related to the estimation of reversible transition matrices, e.g. situa-

tions in which *a priori* information about stationary probabilities of the chain is available.

Uncertainty quantification for reversible Markov state models is difficult. The posterior distribution of reversible transition matrices is supported on a high dimensional domain having a complicated shape. Standard Monte Carlo methods cannot sample efficiently from the posterior. The choice of prior is also problematic, the prior needs to ensure that the posterior mass is concentrated around the true model even if the likelihood is flat in some directions, i.e. only few observations of the rare event are available in the data. Previous approaches to posterior sampling of reversible transition matrices [63, 73] did not provide a prior that was able to compute reliably uncertainties in situations in which only limited sampling of a metastable system was available. Furthermore the algorithms in [63, 73] suffered from long autocorrelation times. Methods that use perturbation approaches are problematic because they use linear approximations to propagate errors. But, linear approximations are only valid if posteriors are sharply peaked which is unlikely to hold for metastable systems in the finite data regime. Computing statistical errors is crucial for any data-driven approach since they provide the only means to assess when sufficient data has been used to estimate the model.

Methods for a reliable quantitative prediction of uncertainties for reversible MSMs are developed. A prior distribution ensuring that the posterior mass is concentrated in the vicinity of the true model for simulation data containing rare events is constructed. An efficient Markov chain Monte Carlo (MCMC) algorithm for sampling from the posterior of reversible transition matrices is developed. Autocorrelation times for the proposed algorithm are much smaller compared to a previous approach in [73]. The algorithm is computationally efficient and can be applied to Markov models with a large number of states.

Kinetic quantities cannot be estimated using enhanced sampling methods, but they can provide reliable estimates of stationary probabilities. The detailed balance equation establishes a relation between stationary probabilities and transition probabilities. This connection offers the possibility to combine information from enhanced sampling simulations and standard molecular dynamics simulations. Furthermore detailed balance relates the forward and the backward direction of a reversible Markov process. For metastable systems it is often much more efficient to sample only one of the two directions. Enhanced sampling data and standard simulations can be combined using the dTRAM method but the resulting optimization problems are large and existing algorithms that are based on a self-consistent iteration can take a long time to converge.

The algorithmic strategy for the solution of the reversible MLE problem is extended to also cover the dTRAM problem. The proposed algorithm can solve the dTRAM problem orders of magnitude faster

than the self-consistent iteration proposed in [123]. The algorithm can efficiently compute the dTRAM estimator for enhanced sampling simulations at many different biasing conditions on large state spaces. Similar to the reversible MLE problem a number of problems related to dTRAM can be solved efficiently.

A systematic approach for the estimation of transition probabilities for rare event systems using reversible MSMs is outlined. The method combines enhanced sampling simulations and short relaxation trajectories to efficiently estimate probabilities for rare transitions. The method uses the detailed balance condition to enable the estimation of a Markov state model in situations in which *only* one direction of the interesting process has been sampled *and* stationary probabilities for the states of the MSM are available from enhanced sampling simulations. It is demonstrated that the method can be used to obtain reliable estimates orders of magnitude before a single rare event would have been observed on average.

Reversible Markov state models are introduced as the discrete space, finite data counterpart to the transfer operator for conformation dynamics. A systematic approach to incorporate reversibility into statistical estimation and uncertainty quantification is developed. Reversibility of the discrete space transition matrix is not enforced for the sake of coherence with the continuous theory alone. It is demonstrated that reversibility can be a key asset in the efficient estimation of probabilities for rare event transitions which previously was a computational bottleneck for the successful application of Markov state models to conformation dynamics.

OUTLINE

In Chapter 2 we introduce the transfer operator as a bounded linear operator on a suitable space of functions and give a short introduction to molecular dynamics. We introduce Markov state models as the discrete space counterpart to transfer operators and discuss maximum likelihood estimation of transition matrices. A brief introduction to Markov chain Monte Carlo methods, umbrella sampling, and statistical reweighting is also given.

Chapter 3 discusses reversible maximum likelihood estimation. A duality argument is used to transform the reversible maximum likelihood problem into a convex-concave program. An efficient algorithm for the solution based on a primal-dual interior-point method for variational inequalities is outlined. It is shown that the proposed algorithm can significantly speed up the estimation compared to a commonly employed fixed-point iteration. The fixed-point iteration is very sensitive to the data that is used in the estimation procedure in contrast to the proposed algorithm. A new algorithm for the solution of the dTRAM problem is presented. The outlined algorithm is orders of magnitude

faster than a previously proposed fixed-point iteration. A decomposition method based on the Schur complement is used to solve the large system of coupled nonlinear equations so that the dTRAM problem can be efficiently solved even in situations in which data at many different biasing conditions is used.

In Chapter 4 new methods for uncertainty quantification of reversible Markov state models are developed. Uncertainties are computed using a Monte Carlo approach. Transition matrices are sampled from the posterior distribution and used to compute sample statistics of observables of interest. It is shown that it is necessary to choose a prior distribution that concentrates the posterior mass around the true model if data containing rare events is used. A prior distribution on the set of reversible transition matrices is constructed and a Gibbs sampling algorithm to generate samples from the posterior is developed. A similar Gibbs sampling algorithm is used to sample the posterior if the stationary vector of the Markov chain is known *a priori*. An extension to situations in which the stationary vector is known with a prescribed uncertainty is developed. The method makes it possible to combine enhanced sampling simulations and equilibrium molecular dynamics simulations into an estimate of a reversible Markov state model. In contrast to previous methods the developed algorithm uses posterior adapted proposal distributions for each Gibbs sampling update and has small autocorrelation times for relevant observables, so that the algorithm is efficient and can be used for Markov state models with a large number of states.

Chapter 5 outlines a systematic approach for the estimation of transition probabilities for systems with rare events using reversible MSMs. Using algorithms for the estimation (Chapter 3) and uncertainty quantification (Chapter 4) it is demonstrated that a combination of enhanced sampling simulations and short relaxation trajectories can be used to efficiently estimate probabilities for rare transitions. The method is applied to a reversible Markov chain with few states, a Markov state model for a metastable diffusion process in a doublewell potential, a Markov state model for the conformation dynamics of a small peptide, and a Markov state model for a non-Markovian projection of a discrete state space dynamics modeling the diffusive motion of a neuronal vesicle that can attach to a cell membrane. It is shown that the method can be used to obtain reliable estimates orders of magnitude before a single rare event would have been observed on average.

THEORY

Ideally, we would like to study the molecular system of interest under biological conditions in atomistic detail. In most cases it is infeasible to include the complex cellular environment into the simulation. Instead we study the molecule at room temperature in a 'box' of constant volume. It is also necessary to explicitly simulate a fixed number of water molecules surrounding the molecule to correctly capture solvent effects which can be important for conformational processes [20] [1].

From a perspective of statistical mechanics we study the canonical ensemble for the simulated system. We are not only interested in computing equilibrium expectations from equilibrium probabilities of the ensemble. But, we also want to study the equilibrium dynamics (thermal fluctuations) of the system. The equilibrium dynamics can be modeled by a stochastic process, which is a stationary Markov process. The Markov process is time-reversible, i.e. all n-point probabilities for the forward and backward in time process in equilibrium are equal.

In molecular dynamics (MD) the fastest motions (bond vibrations) happen on timescales of femtoseconds ($10^{-15}s$) while conformational changes occur on microsecond ($10^{-6}s$) or milliseconds ($10^{-3}s$). Integrators for MD allow stepsizes of at most some femtoseconds so that conformation dynamics is not accessible via long trajectories due to its rare event character.

Long trajectories have to be replaced by ensembles of short trajectories and conformation dynamics has to be studied in terms of the evolution of ensembles. The evolution of ensembles can be studied via a transfer operator and conformations can be identified as sets in the space of protein configurations which are almost invariant under the action of this operator. This leads to a hybrid approach, where a projection of the high-dimensional transfer operator is estimated from short trajectories giving rise to a transition matrix and interesting properties of the operator are computed from the matrix.

As a result of reversibility the transfer operator fulfills a detailed balance condition. The operator can be regarded as a self-adjoint operator on a weighted L^2-space with weights given by the invariant measure of the process. A statistical estimator for transition probabilities based on relative frequencies of transitions observed in a finite ensemble of trajectories will however result in a matrix which does not fulfill detailed balance, so that an important property of the transfer operator is lost. A reversible transition matrix can be obtained via a probabilistic inter-

---

1 For reasons of computational efficiency, water molecules are usually not resolved in atomistic detail. Instead, most common water models are rigid, i.e. the relative positions of oxygen and hydrogen atoms in the water molecule are fixed.

pretation of the transfer operator leading to a constrained maximum likelihood estimation problem. An efficient method for the solution of this problem is developed in Chapter 3.

2.1   MOLECULAR DYNAMICS

In molecular dynamics the state space X is the phase space $\Omega_q \times \Omega_p \subseteq \mathbb{R}^{2d}$ of coordinates $q \in \mathbb{R}^d$ and momenta $p \in \mathbb{R}^d$. The dimension d is $d = 3N$ for a system with N particles. Since we are interested in systems with a fixed number of particles at constant temperature and constant volume the invariant measure of the process is the canonical distribution

$$\pi(dx) = \frac{1}{Z} \exp(-\beta H(x)) dx. \tag{2.1}$$

The parameter β is the inverse temperature and the constant Z ensuring correct normalization is the partition function

$$Z = \int_X dx \exp(-\beta H(x)). \tag{2.2}$$

The real valued function $H(x)$ is the Hamiltonian, specifying the energy of a state $x = (q, p)$

$$H(x) = \frac{1}{2} p^T M^{-1} p + V(q). \tag{2.3}$$

The matrix M in the kinetic energy term is a diagonal matrix containing the particle masses and $V(q)$ is the molecular potential. Since H contains no mixed p, q-terms the density corresponding to the canonical measure factorizes

$$\pi(x) = \underbrace{\frac{1}{Z_p} \exp(-\frac{\beta}{2} p^T M^{-1} p)}_{\pi_p(p)} \underbrace{\frac{1}{Z_q} \exp(-\beta V(q))}_{\pi_q(q)}. \tag{2.4}$$

In molecular dynamics the potential function has two parts. One for short-range and one for long-range interactions

$$V(q) = V_{\text{short}}(q) + V_{\text{long}}(q). \tag{2.5}$$

Short-range or bonded interactions describe harmonic bond vibrations, bending of bond-angles, rotation about dihedral angles, and deviations from the planar geometry of certain groups [68]

$$V_{\text{short}}(q) = \sum_{\text{bonds}} \frac{k_d}{2} (d(q) - d_0)^2 + \sum_{\text{angles}} \frac{k_\theta}{2} (\theta(q) - \theta_0)^2 +$$
$$\sum_{\text{dihedrals}} \frac{k_\phi}{2} (1 + \cos(n\phi(q) - \phi_0)) + \tag{2.6}$$
$$\sum_{\text{impropers}} \frac{k_\psi}{2} (\psi(q) - \psi_0)^2.$$

Long-range or non-bonded interactions model van der Waals and electrostatic interactions of particles separated by at least three bonds [68]

$$V_{\text{long}}(q) = \sum_{\substack{\text{non-bonded} \\ \text{pairs (i, j)}}} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{q_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{q_{ij}}\right)^{6}\right] + \sum_{\substack{\text{non-bonded} \\ \text{pairs (i, j)}}} \frac{c_i c_j}{4\pi\epsilon_0 q_{ij}}. \tag{2.7}$$

A detailed discussion of the individual terms and parameters in (2.6) and (2.7) can be found in [68].

It should be noted that the molecular potential is a simplified, empirical model facilitating a classical description of the quantum mechanical interactions between nuclei and electrons in terms of the nuclei positions alone.

2.1.1 *Langevin dynamics*

Equations of motion for the molecular system at constant temperature and constant volume can be derived using the Mori-Zwanzig projection operator formalism [128]. The resulting equations of motion are closed, i.e. all forces can be computed without information about the positions and momenta of the particles in the environment. The interaction with the environment is described via effective friction and stochastic forces modeling the effect of random collisions with unresolved particles in the environment. The resulting equations of motion are the Langevin equations

$$\begin{aligned} \dot{q} &= M^{-1}p, \\ \dot{p} &= -\nabla_q V(q) - \gamma M^{-1}p + \sigma\dot{W}. \end{aligned} \tag{2.8}$$

Equation (2.8) is a stochastic differential equations (SDE). The constant γ is the damping factor for the friction force, σ is the noise intensity of the stochastic force modeled by a d-dimensional Wiener process W_t with mean $\langle W_t \rangle = 0$ and correlation $\langle W_t W_s \rangle = \delta(t - s)$. The solution to the Langevin equation (2.8) is a time-homogeneous Markov process.

If one of the following two conditions is satisfied

(P) Periodic system: Position space $\Omega_q \subset \mathbb{R}^d$ is periodic and the potential $V : \Omega_q \to \mathbb{R}$ is smooth

(B) Bounded system: Position space is $\Omega_q = \mathbb{R}^d$, the potential V is smooth, bounded from below and $V(q)$ is growing at infinity as $\|q\|^{2l}$ for some positive integer l. Such systems are called bounded, since the energy surfaces $\{x | H(x) = E\}$ are bounded subsets of X [95].

then the solution process has a unique invariant measure given by the canonical measure π in equation (2.1). Inverse temperature β, noise intensity σ, and damping factor γ satisfy the following equation [8]

$$\sigma^2 = \frac{2\gamma}{\beta}. \tag{2.9}$$

2.1.2   *Brownian dynamics*

In the high friction limit $\gamma \to \infty$ the Langevin equation can be simplified to yield Smoluchowski or Brownian dynamics [8]

$$\dot{q} = -\nabla V(q) + \sqrt{2\beta^{-1}}\dot{W}. \tag{2.10}$$

The solution of (2.10) is a time-homogeneous Markov process. If condition (P) or (B) are satisfied then the canonical density π_q is the unique invariant density for Brownian dynamics.

2.2   TRANSFER OPERATOR

This section reviews the transfer operator approach to molecular dynamics. The presentation closely follows [95].

Let $(X, \mathcal{A}, \mu)$ be a probability space, where $X \subseteq \mathbb{R}^d$ is the state space, $\mathcal{A}$ the Borel σ-algebra on X and μ is a probability measure on $\mathcal{A}$. If the measure μ is absolutely continuous with respect to the Lebesgue measure dx on X then it has a density such that

$$\mu(A) = \int_A dx\, \mu(x), \quad \forall A \in \mathcal{A}.$$

This slight abuse of notation should not be too confusing. It will be clear from the context if μ is a density or a measure.

For a given measure μ we denote by $L^p(\mu)$ the usual L^p space of (real-valued) functions on X. The space $L^2(\mu)$ is a Hilbert space with scalar product

$$\langle u, v \rangle_\mu = \int \mu(dx) u(x) v(x).$$

A time-homogeneous Markov process $(X_t)_{t \in I}$ in continuous or discrete time, $I = \mathbb{R}_0^+$, or $I = \mathbb{N}_0$ can be defined in terms of a (stochastic) transition function $p : I \times X \times \mathcal{A} \to [0, 1]$ [2] that satisfies the following conditions

1. $p(t, x, \cdot)$ is a probability measure on $\mathcal{A}$ for every $t \in I$, $x \in X$ and $p(0, x, X \setminus \{x\}) = 0$ for every $x \in X$, i.e. $p(0, x, \cdot)$ is the Dirac measure $\delta_x(\cdot)$ for all $x \in X$.

---

2 We assume that X_t is measurable and nonsingular with respect to the base measure μ, i.e. $\mu(X_t^{-1}(A)) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) = 0$.

2. $p(t, \cdot, A)$ is measurable for every $t \in I$, $A \in \mathcal{A}$

3. $p(\cdot, x, A)$ satisfies the Chapman-Kolmogorov equation

$$p(t + s, x, A) = \int p(t, x, dy)p(s, y, A) \tag{2.11}$$

for all $t, s \in I$, $x \in X$, and $A \in \mathcal{A}$.

The transition function specifies the conditional probability for finding the process $(X_t)_{t \in I}$ after a time t in the set A given that it was started in x

$$\mathbb{P}[X_{s+t} \in A | X_s = x] = p(t, x, A), \quad \forall s, t \in I, \ A \in \mathcal{A}. \tag{2.12}$$

Equation (2.12) is invariant with respect to the starting time s since the process $(X_t)_{t \in I}$ is time-homogeneous.

The transition function describes the transport of probability measures. Let ν_0 be a probability measure on $(X, \mathcal{A})$ then

$$\nu_t(A) = \int_X \nu_0(dx) \, p(t, x, A) \tag{2.13}$$

is also a probability measure on $(X, \mathcal{A})$. [3]

The transition function induces a linear operator $P_t : L^1(\mu) \to L^1(\mu)$. For p absolutely continuous with respect to the base measure μ it is defined as

$$(P_t u)(y) = \int_X \mu(dx)u(x)p(t, x, y), \quad \forall u \in L^1(\mu). \tag{2.14}$$

In general p will not be absolutely continuous with respect to μ.

The backward operator $T_t : L^\infty(\mu) \to L^\infty(\mu)$ is defined as

$$(T_t v)(x) = \int_X p(t, x, dy)v(y), \quad \forall v \in L^\infty(\mu). \tag{2.15}$$

The integral in (2.15) is the conditional expectation value of the random variable $v(X_t)$ for the process $(X_t)_{t \in I}$ starting in x at $t = 0$,

$$(T_t v)(x) = \mathbb{E}[v(X_t)|X_0 = x]. \tag{2.16}$$

Under mild regularity conditions the operator T_t is a well defined operator on $L^\infty(\mu)$ and the transfer P_t operator can be defined as the adjoint of the backward operator T_t [69].

The operator P_t describes the time-evolution of probability densities and T_t describes the time-evolution of expectation values.

In the time-continuous setting, $I = \mathbb{R}_0^+$, the set $(P_t)_{t \in I}$ forms a strongly continuous semigroup of transfer operators [4], i.e.

$$\lim_{t \to 0} P_t u = u, \quad P_{t+s} u = P_t P_s u, \quad \forall u \in L^1_\mu(X). \tag{2.17}$$

---

3 The integral is well defined since the measurable function $p(t, \cdot, A)$ is bounded $\forall A \in \mathcal{A}$.

4 (2.17) is a result of $p(0, x, X \setminus \{x\}) = 0$ for all $x \in X$ and the Chapman-Kolmogorov equation (2.11)

2.2.1 *Invariant measures*

Invariant measures are invariant under the action of the transfer function, i.e.

$$\pi(A) = \int_X \pi(dx)p(t,x,A), \quad \forall t \in I, \, A \in \mathcal{A}. \tag{2.18}$$

If we chose $(X, \mathcal{A}, \pi)$ as the probability space for the transfer operator then the characteristic function of X is an invariant density of P_t,

$$P_t \chi_X = \chi_X, \quad \forall t \in I. \tag{2.19}$$

2.2.2 *Detailed balance and reversibility*

If the detailed balance equation holds for the transition function

$$\int_A \pi(dx)p(t,x,B) = \int_B \pi(dx)p(t,x,A), \quad \forall A, B \in \mathcal{A} \tag{2.20}$$

then the measure π is the invariant measure for the process and the transfer operator is a self-adjoint operator on $L^2(\pi)$ [5], i.e.

$$\langle v, P_t u \rangle_\pi = \langle P_t v, u \rangle_\pi, \quad \forall u, v \in L^2(\pi). \tag{2.21}$$

If the detailed balance equation (2.20) holds for a stationary Markov process $(X_t)_{t \in I}$, i.e. $(X_t)_{t \in I}$ is a time-homogeneous process with initial condition distributed according to its invariant measure, then all n-point probabilities for the forward and the time-reversed process are equal, i.e. for all $t_1 \leq \cdots \leq t_n \in I$ and for all $A_1, \ldots, A_n \in \mathcal{A}$

$$\mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = \mathbb{P}(X_{s_n} \in A_1, \ldots, X_{s_1} \in A_n). \tag{2.22}$$

Time-points for the time-reversed process on the right hand side of (2.22) are given by $s_k = t_n - (t_{n-k+1} - t_1)$ [78].

2.2.3 *Probabilistic interpretation*

We can extend the definition for the conditional probability in (2.12) to the case $X_t \in B$ given that $X_0 \in A$,

$$p(t, A, B) = \mathbb{P}[X_t \in B | X_0 \in A] = \frac{1}{\mu(A)} \int_A \mu(dx)p(t,x,B). \tag{2.23}$$

Equation (2.23) can be interpreted as the following two step experiment [95]

1. Preparation: Select from an ensemble distributed according to the base measure μ all systems in set A.

---

5 The inclusion $L^p(\mu) \subset L^1(\mu)$ for any $p \in [1, \infty]$ follows from the Hölder inequality since μ is a finite measure, i.e $\mu(X) < \infty$.

2. Measurement: Determine the relative frequency of selected systems ending up in the set B after time t.

The probability $p(t, A, B)$ in (2.23) will in general depend on the chosen base measure μ. Since we are interested in equilibrium fluctuations we assume from now on that the base measure is chosen as the invariant measure of the process π, i.e. the initial ensemble in the experiment is in equilibrium.

In terms of the transfer operator equation (2.23) can be expressed as follows

$$p(t, A, B) = \frac{\langle \chi_B, P_t \chi_A \rangle_\mu}{\langle \chi_A, \chi_A \rangle_\mu}. \tag{2.24}$$

2.2.4 *Almost invariant sets*

Almost invariant sets can now be characterized using the interpretation of the set-to-set transition probability (2.23) via the above two step experiment for the equilibrium ensemble.

A set C is almost invariant if the relative frequency of systems (in equilibrium) staying in C after some characteristic time-span t is close to 1:

$$C \text{ almost invariant} \quad \Leftrightarrow \quad p(t, C, C) \approx 1. \tag{2.25}$$

In terms of the transfer operator almost invariance of a set C means that

$$P_t \chi_C \approx \chi_C \tag{2.26}$$

where χ_C is the characteristic function of the set C.

If the set C is invariant then $P_t \chi_C = \chi_C$, i.e. χ_C is an eigenvector corresponding to the eigenvalue $\lambda = 1$. According to (2.26) χ_C is an approximate eigenvector for an eigenvalue close to the eigenvalue $\lambda = 1$.

The central idea of the transfer operator approach is: Identification of invariant sets via eigenvectors corresponding to dominant eigenvalues, i.e. eigenvalues close to the Perron eigenvalue $\lambda = 1$. For a successful identification strategy we require the following two conditions to hold for the transfer operator:

(C1) P_t is asymptotically stable: $(P_t)^n u \to \chi_X$ as $n \to \infty$ for every density $u \in L^1(\pi)$, i.e. the eigenvalue $\lambda = 1$ is simple and dominant.

(C2) The essential spectrum of P_t, i.e. the spectrum excluding eigenvalues of finite multiplicity, is strictly bounded away from the unit circle $|\lambda| = 1$.

Uniqueness of the invariant measure (simple eigenvalue $\lambda = 1$) is important since almost invariance is defined with respect to a selected

measure which is assumed to uniquely represent the equilibrium ensemble of the system of interest. Dominance of the eigenvalue $\lambda = 1$ excludes other eigenvalues with $|\lambda| = 1$ so that any initial ensemble will converge to the equilibrium ensemble. The condition for the essential spectrum ensures that the dominant eigenvalues of P_t can be used in a computational approach to the identification of almost invariant sets [95].

2.2.5 *Transfer operator for molecular dynamics*

This section reviews properties of the transfer operator for molecular dynamics described by Langevin or Brownian dynamics. It is assumed that either condition (P) or (B) is satisfied for the molecular potential.

The solution to the Langevin equation (2.8) is a time-homogeneous Markov process $(X_t)_{t \in I} = (q_t, p_t)_{t \in I}$ with invariant measure π given by the canonical measure in equation (2.1). The transition function p for the Langevin process is absolutely continuous with respect to the invariant measure for all $t > 0$ and it induces a semi-group of transfer operators $(P_t)_{t \in I}$ [8].

The corresponding transition kernel and the corresponding invariant density fulfill the following extended detailed balance equation [96]

$$\pi(x)p(t,x,y) = \pi(Ry)p(t,Ry,Rx), \quad \forall t > 0, \ x, y \in X \qquad (2.27)$$

with momentum reversal operator $Rx = R(q,p) = (q,-p)$.

In most cases one is interested in the dynamics of molecular configurations, i.e. the induced process $(q_t)_{t \in I}$. The momenta p_t are usually ignored.

In [8] it was shown that for the Langevin equation (2.8) the induced process $(q_t)_{t \in I}$ is a Markov process with absolutely continuous transition function p satisfying the detailed balance equation

$$\pi_q(q_1)p(t,q_1,q_2) = \pi_q(q_2)p(t,q_2,q_1), \quad \forall t > 0, \ q_1, q_2 \in \Omega_q. \quad (2.28)$$

As a result the corresponding spatial transition operator S_t is a self-adjoint operator on $L^2(\pi_q)$. If conditions (P) or (B) are fulfilled for the potential then S_t possesses the necessary spectral properties (C1) and (C2) for identification of almost invariant sets [8].

The transition kernel for the spatial dynamics does not obey the Chapman-Kolmogorov equation

$$p(t_1 + t_2, x, y) \neq \int dz p(t_1, x, z)p(t_2, z, y). \qquad (2.29)$$

so that spatial transfer operators do not form a semigroup, $S_{t+s} \neq S_t S_s$.

Observing that the relaxation of the momenta is significantly faster than that of the positions it is argued in [9] that the position dynamics is "almost Markovian" on timescales larger than a suitable lagtime τ

$$p(t + s, x, y) \approx \int dz p(t, x, z)p(s, z, y), \quad \forall t, s \geq \tau. \qquad (2.30)$$

As a result long-time predictions can be computed from the spatial transfer operator S_t with a time resolution coarser than the lagtime

$$S_{nt} \approx (S_t)^n, \quad \forall t > \tau, \, n \in \mathbb{N}. \tag{2.31}$$

Brownian dynamics, i.e. the solution to equation (2.10), is a time-homogeneous Markov process $(q_t)_{t \in I}$. The transition function for the process is absolutely continuous with respect to canonical density π_q from (2.4). The transition function satisfies the Chapman-Kolmogorov equation (2.11) and induces a strongly continuous semigroup of (spatial) transfer operators $(S_t)_{t \in I}$.

The transition kernel and the fulfills the detailed balance equation for the transition function so that the transfer operator S_t is a self-adjoint operator on $L^2(\pi_q)$.

If the potential fulfills either condition (P) or (B) then the transfer operator for Brownian dynamics possesses again the desirable spectral properties (C1) and (C2) for the identification of almost invariant sets [95].

2.3 MARKOV STATE MODELS

The two step experiment outlined in Section 2.2.3 suggests the computation of $p(t, A, B)$ in terms of relative frequencies

$$p(t, A, B) = \lim_{N_A \to \infty} \frac{N_{AB}(t)}{N_A}. \tag{2.32}$$

Where N_A is the total number of systems in the (equilibrium) ensemble starting in A and $N_{AB}(t)$ is the number of such systems that end up in B after time t.

Previously, molecular conformations were identified as metastable subsets of the state space (position space). In Section 2.2.4 it was outlined how such sets can be identified from eigenvectors corresponding to the dominant eigenvalues of the transfer operator P_t. Unfortunately P_t is almost never available directly.

The idea central to Markov state models (MSM) is to obtain a finite dimensional approximation of P_t in a way that allows to identify matrix elements of the discretized operator with transition probabilities between sets. These transition probabilities can then be estimated from simulation data.

Let $S_1, \ldots, S_d$ be a partition of the state space X into sets with

$$X = \bigcup_{i=1}^{d} S_i, \quad S_i \bigcap S_j = \emptyset, \, i \neq j. \tag{2.33}$$

Then a corresponding space of ansatz functions for a Galerkin discretization of P_t is

$$V = \text{span}\{\chi_1, \ldots, \chi_d\}. \tag{2.34}$$

The function χ_k is the characteristic function of the set S_k. The Galerkin projection $\Pi : L^2(\pi) \to V$ is

$$\Pi u = \sum_{k=1}^{d} \frac{\langle \chi_k, u \rangle_\pi}{\langle \chi_k, \chi_k \rangle_\pi} \chi_k = \sum_k u_k \chi_k, \quad \forall u \in L^2(\pi). \tag{2.35}$$

The Galerkin projection of the transfer operator P_t is

$$\begin{aligned} \Pi P_t \Pi u &= \sum_{i,j} \chi_i \frac{\langle \chi_i, P_t \chi_j \rangle_\pi}{\langle \chi_i, \chi_i \rangle_\pi} u_j = \sum_{i,j} \chi_i \frac{\langle P_t \chi_i, \chi_j \rangle_\pi}{\langle \chi_i, \chi_i \rangle_\pi} u_j \\ &= \sum_{i,j} \chi_i p_{ij} u_j. \end{aligned} \tag{2.36}$$

For the second equality we have used that P_t is self-adjoint on $L^2(\pi)$. The matrix element p_{ij} is the set-to-set transition probability from equation (2.24)

$$p_{ij} = p(t, S_i, S_j) = \frac{\langle P_t \chi_i, \chi_j \rangle_\pi}{\langle \chi_i, \chi_i \rangle_\pi}. \tag{2.37}$$

The transition matrix $P = (p_{ij})$ inherits the following four important properties from P_t [95]:

(M1) The matrix P is (row-)stochastic, i.e. $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$ for all $i, j = 1, \ldots, d$. As a result all eigenvalues λ satisfy $|\lambda| \leq 1$.

(M2) The row vector $\pi = (\pi_1, \ldots, \pi_d)$ with $\pi_i = \int_{S_i} \pi(dx)$ is a left invariant vector corresponding to the eigenvalue $\lambda = 1$, i.e. $\pi^T P = \pi^T$.

(M3) The vector π satisfies the detailed balance equation for the transition matrix P, i.e. $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i, j = 1, \ldots, d$. As a result, the matrix P is symmetric with respect to the π-weighted scalar product $\langle u, v \rangle_\pi = \sum_i \pi_i u_i v_i$, i.e. $\langle u, Pv \rangle_\pi = \langle Pu, v \rangle_\pi$ for all $u, v \in \mathbb{R}^d$ so that all eigenvalues of the matrix P are real.

(M4) If the transfer operator P_t satisfies (C1), then the matrix P is irreducible and aperiodic, i.e. the eigenvalue $\lambda = 1$ is simple and dominant so that the vector π is the unique stationary vector.

The projection Π destroys the semigroup property of (P_t), i.e.

$$\Pi P_{t+s} \Pi \neq (\Pi P_t \Pi)(\Pi P_s \Pi) \tag{2.38}$$

so that the n-step transition probabilities between sets can not be described by powers of the transition matrix P

$$p(nt, S_i, S_j) \neq (P^n)_{ij}. \tag{2.39}$$

One can however show that the semi-group property can be approximately recovered for a suitable choice of lagtime τ, if the dominant

eigenvectors of P_t can be approximated by linear combinations of the chosen ansatz functions. This implies that the chosen sets can approximately resolve the slow processes with a time resolution coarse than the lagtime τ [84, 91].

A complete Markov state model is specified by sets $(S_i)_{i=1}^d$ for the discretization, an appropriate lagtime τ ensuring approximate Markovianity of the projected operator, and the transition matrix $P = \Pi P_\tau \Pi$ estimated from data

$$\text{MSM} = \{(S_i)_{i=1}^d, \tau, P\}. \tag{2.40}$$

A Galerkin projection using a more general class of ansatz functions can also be performed. Slow processes can then be computed using variational methods [74, 79, 80]. But, the resulting matrix elements cannot be interpreted as probabilities and have to be estimated via equilibrium correlation functions.

2.4 MAXIMUM LIKELIHOOD ESTIMATION

Previously, we have seen that the matrix elements resulting from a discretization of the transfer operator with characteristic functions can be interpreted as conditional set-to-set transition probabilities which can be estimated from simulation data.

Once the set-based discretization has been chosen the high-dimensional time-series of configurations generated by the molecular dynamics simulation (q_t) is mapped to a time-series of integer labels (X_t) via the simple membership rule

$$q_t \in S_i \Rightarrow X_t = i. \tag{2.41}$$

We assume that at a suitable lagtime τ the resulting jump process $(X_k)_{k \in \mathbb{N}_0}$ with $X_k = X_{k\tau}$ is a time-homogeneous Markov chain on a finite state space, i.e.

$$\mathbb{P}(X_{k+1}|X_k, \ldots, X_1, X_0) = \mathbb{P}(X_{k+1}|X_k), \quad \forall k \in \mathbb{N}_0. \tag{2.42}$$

The matrix elements for the MSM are then given by the conditional jump probabilities of the chain

$$p_{ij} = \mathbb{P}(X_{k+1} = j | X_k = i), \quad \forall i, j = 1, \ldots, d. \tag{2.43}$$

The probabilities p_{ij} have to be estimated from a finite realization of the chain $X_0 = i_0, \ldots, X_N = i_N$.

An estimator $\hat{p}_{ij} = \hat{p}_{ij}(X_0, \ldots, X_N)$ for the matrix elements p_{ij} can be found by maximizing a likelihood function. The resulting estimator is called the maximum likelihood estimator (MLE). Under certain regularity conditions [34] the MLE has the following desirable asymptotic properties:

1. Consistency, i.e. the MLE converges to the true transition probabilities

2. Normality, i.e. the asymptotic distribution of the MLE is a normal distribution with mean given by the true transition probabilities

3. Efficiency, i.e. asymptotically the MLE is the best unbiased estimator for the transition probabilities - that is it has minimum variance among all unbiased estimators

Using the Markov property of the process $(X_k)_{k \in \mathbb{N}_0}$ the likelihood of observing a finite length random sequence $X_0, \ldots, X_N$ for a given transition matrix $P = (p_{ij})$ can be expressed as

$$\mathbb{P}(X_0, X_1, \ldots, X_N | P) = \prod_{k=0}^{N-1} \mathbb{P}(X_{k+1} | X_k) \mathbb{P}(X_0). \qquad (2.44)$$

Denote by $c_{ij} = c_{ij}(X_0, \ldots, X_N)$ the number of transitions in the sequence $X_0, \ldots, X_N$ going from state i to state j. Using the transition counts c_{ij} we can rearrange the product in (2.44)

$$\mathbb{P}(X_0, X_1, \ldots, X_N | P) = \mathbb{P}(X_0, C | P) = \prod_{i,j} p_{ij}^{c_{ij}} \mathbb{P}(X_0). \qquad (2.45)$$

Equation (2.45) shows that the likelihood of the sequence $X_0, \ldots, X_N$ is determined by the probability distribution of the initial state $\mathbb{P}(X_0)$ and the matrix of transition counts $C = (c_{ij})$. The transition counts c_{ij} are a minimal sufficient statistic for the transition probabilities p_{ij} if the distribution of the initial state is independent of the p_{ij} which is assumed from now on [2]. All information about the desired parameters p_{ij} contained in $X_0, \ldots, X_N$ can be minimally encoded into C.

The maximum likelihood estimator $\hat{P}$ can be found by maximizing the following log-likelihood over the set of all stochastic matrices

$$L(C | P) = \sum_{ij} c_{ij} \log p_{ij}. \qquad (2.46)$$

Using the method of Lagrange multipliers one can show that the MLE is simply given by $\hat{P} = (\hat{p}_{ij})$ [2] with

$$\hat{p}_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}. \qquad (2.47)$$

The above argument can be extended to the case of an ensemble of M independent finite length chains $\{(X_{1,k})_{k=1}^{N_1}, \ldots, (X_{M,k})_{k=1}^{N_M}\}$. The transition counts are then given by the transition counts accumulated over all chains in the ensemble,

$$c_{ij} = \sum_{l=1}^{M} c_{l,ij}.$$

The integer $c_{l,ij}$ denotes the number of transitions from state i to state j in the l-th chain $(X_{l,k})_{k=1}^{N_l}$. The resulting MLE is again given by equation (2.47).

The estimator in (2.47) is exact in the limit of an infinitely long realization $(N \to \infty)$ starting from a fixed initial state $X_0 = i_0$ *or* in the limit of an infinite ensemble of independent finite length chains $(M \to \infty, N \in \mathbb{N})$ under the condition that the probability to start in state i is positive for all $i = 1, \ldots, d$ [2].

For a finite observation of the chain the estimator in (2.47) is not reversible even if the chain $(X_k)_{k \in \mathbb{N}_0}$ is reversible. An important property (M3) that holds asymptotically is violated when using the estimator in equation (2.47) for any finite amount of data. A solution to the MLE problem that ensures reversibility is presented in Chapter 3.

2.5 MARKOV CHAIN MONTE CARLO

The objective here is to generate samples from a target density $\pi(x)$, $x \in X \subseteq \mathbb{R}^d$. To simplify the presentation we assume that X is a connected set and that $\pi(x) > 0$ for all $x \in X$. The target density π is possibly non-normalized but is integrable.

2.5.1 *Rejection sampling*

This section closely follows [25]. Rejection sampling relies on the ability to generate samples from a proposal density $\rho(x)$ with the property that for a known constant $c > 0$

$$\pi(x) \leq c\rho(x), \quad \forall x \in X. \tag{2.48}$$

The proposal density $\rho(x)$ is possibly non-normalized but is integrable. Samples from the target density π are then generated via proposals sampled according to ρ. The proposal x is accepted if

$$u \leq \frac{\pi(x)}{c\rho(x)} \tag{2.49}$$

for a random u which is uniformly distributed in $[0, 1]$, cf. Algorithm 1. Rejection sampling is efficient whenever sampling from ρ is efficient and the average number of rejections is small.

The following property of densities is central to the idea of rejection sampling. Random points (x, y) with $x \sim \rho(x)$, $y = u\rho(x)$ and u uniformly distributed in $[0, 1]$ are uniformly distributed in the hypograph of ρ, i.e. the area under the graph of ρ,

$$\text{hyp}\, \rho = \{(x, y) | 0 \leq y \leq \rho(x)\}. \tag{2.50}$$

Vice versa, if (x, y) is uniformly distributed over $\text{hyp}\, \rho$, then x is distributed according to ρ. The same applies to $c\rho$ for any $c > 0$.

Algorithm 1 : Rejection method

Input : ρ, π, c
Output : x
while *True* **do**
    Generate $x \sim \rho(x)$ and u uniformly in $[0,1]$
    if $u \leq \frac{\pi(x)}{c\rho(x)}$ **then**
        Return x
    end
end

If ρ is a valid proposal density for π, i.e (2.48) holds for some $c > 0$, then hyp π is contained in hyp $c\rho$. In the rejection method we propose points uniformly in hyp $c\rho$ accepting only points that lie in hyp π to generate x distributed according to π.

The acceptance probability p for the rejection method is inverse proportional to the constant c,

$$p = \mathbb{P}\left[u \leq \frac{\pi(x)}{c\rho(x)}\right] = \frac{1}{c}. \tag{2.51}$$

The probability of accepting a sample after k proposals, i.e. after $k-1$ rejections, is $(1-p)^{k-1}p$. The expected number of proposals for successfully generating a sample is thus given by

$$p\sum_{k=1}^{\infty} k(1-p)^{k-1} = c. \tag{2.52}$$

The optimal value c^* should be as small as possible, so that

$$c^* = \sup_{x \in X} \frac{\pi(x)}{\rho(x)}. \tag{2.53}$$

2.5.2 *Metropolis Hastings algorithm*

The presentation is similar to [17], but the material can be found in almost every textbook on Monte Carlo sampling, e.g. [87]. For the rejection method samples from the proposal density ρ are usually drawn independently. The central idea of the Metropolis Hastings algorithm is to regard the target density π as the invariant density of a Markov chain with unknown transition function p. Knowledge of π, i.e. the ability to evaluate it for a given x, and a proposal generating transition kernel q are then used to generate a chain evolving according to p. The algorithm is constructed such that π and p satisfy the detailed balance equation (2.20) which ensures that the target density π is the invariant density of the chain.

Starting from a valid initial point $x_0 \in X$ the algorithm generates a sequence of points $x_1, \ldots, x_N$ which will asymptotically be distributed

according to π. In the $(i+1)$-th step proposals y are sampled from the proposal kernel $q(x_i, \cdot)$. If the acceptance criterion

$$u \leq \min\left\{1, \frac{\pi(y)q(y, x_i)}{\pi(x_i)q(x_i, y)}\right\} \tag{2.54}$$

is satisfied for a random u which is uniformly distributed in $[0, 1]$ then the chain moves to the proposed point, $x_{i+1} = y$, otherwise the chain remains at its current position, $x_{i+1} = x_i$, cf. Algorithm 2. The algorithm is efficient whenever sampling from q is efficient for all $x \in X$ and the correlation between samples decays fast.

Algorithm 2 : Metropolis-Hastings algorithm

Input : q, π, x_0, N
Output : $x_1, \ldots, x_N$
for $i = 0, \ldots, N-1$ **do**
  Generate $y \sim q(x_i, y)$ and u uniformly in $[0, 1]$
  if $u \leq \min\left\{1, \frac{\pi(y)q(y, x_i)}{\pi(x_i)q(x_i, y)}\right\}$ **then**
    $x_{i+1} = y$
  else
    $x_{i+1} = x_i$
  end
end

Samples $x_1, \ldots, x_N$ generated from the Markov chain with transition function p starting from a valid initial point $x_0 \in X$ will asymptotically be distributed according to π if the chain is π-irreducible and aperiodic [87, Theorem 7.4]. This essentially holds for the Metropolis-Hastings chain if the proposal density satisfies the following positivity condition [87, Corollary 7.5]

$$q(x, y) > 0, \quad \forall x, y \in X. \tag{2.55}$$

It is possible to establish convergence under weaker assumptions on π and q, cf. [87, Corollary 7.7].

If π would already satisfy the detailed balance equation (2.20) for our proposal kernel q we would be done. In all other cases we modify q via an acceptance kernel $0 \leq a(x, y) \leq 1$ to ensure that the detailed balance equation (2.20) holds for the resulting transition function p. Rejecting a proposed move means keeping the current sample x so that the transition function contains a point mass

$$p(x, A) = r_x \delta_x(A) + \int_A dy\, a(x, y)q(x, y). \tag{2.56}$$

The acceptance kernel is given by the Metropolis-Hastings rule

$$a(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{2.57}$$

and the rejection probability r_x is given by

$$r_x = 1 - \int_X dy\, a(x,y)q(x,y). \tag{2.58}$$

Central to the Metropolis-Hastings algorithm is the following idea: If $\pi(x)q(x,y) > \pi(y)q(y,x)$ for some $x,y \in X$ then the chain evolving according to q moves from x to y too often and from y to x too seldom. The acceptance kernel is chosen such that it corrects this imbalance, i.e. the detailed balance equation $\pi(x)a(x,y)q(x,y) = \pi(y)a(y,x)q(y,x)$ holds. Since the q-chain moves from y to x too seldom we always accept the move from y to x, so that $a(y,x) = 1$ and $a(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} < 1$.

If the proposal kernel is symmetric, i.e. $q(x,y) = q(y,x)$, then the acceptance probability (2.57) depends only on π. The symmetric case in which $q(x,y) = q(|x - y|)$ is called a random-walk proposal since the candidate is the current value plus a proposal increment. The most prominent distribution for the random-walk proposal is the Normal distribution $\mathcal{N}(0,\sigma^2)$ with zero mean and variance σ^2.

2.5.3 *Gibbs sampling*

The Gibbs sampler generates samples from the target distribution π via successive updates of entries x_k of a vector $x = (x_1, \dots x_d) \in X$. Entries x_k are sampled in succession from their corresponding conditional distributions

$$\pi_k(x_k) = \pi(x_k | x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_d). \tag{2.59}$$

Once the entry x_k has been updated the next entry x_{k+1} is sampled conditioned on the previously updated coordinates. After a sweep over all entries the algorithm returns a new vector, cf. Algorithm 3. The algorithm is efficient whenever sampling from the set of conditionals $(\pi_k)_{k=1}^d$ is efficient for all $x \in X$ and the correlation between samples decays fast.

Algorithm 3 : Gibbs sampler

> **Input :** π, $x^{(0)}$, N
> **Output :** $x^{(1)}, \dots, x^{(N)}$
> **for** $i = 1, \dots, N$ **do**
> $x = x^{(i)}$
>   **for** $k = 1, \dots, d$ **do**
> $y_k \sim \pi(y_k | y_1, \dots, y_{k-1}, x_{k+1}, \dots, x_d)$
>   **end**
> $x^{(i+1)} = y$
> **end**

The update rule ensures that the resulting sequence of $x_1, \ldots, x_N$ is asymptotically distributed according to π if the following positivity condition holds for all $k = 1, \ldots, d$

$$\pi_k(x_k | x_1, \ldots, x_{k-1}, y_{k+1}, \ldots, y_d) > 0, \quad \forall x, y \in X. \tag{2.60}$$

Convergence can established under weaker assumptions, cf. [87, Corollary 10.12].

2.5.4 *Sampling errors*

Most Monte Carlo methods are used to approximate expectation values with respect to a target density π. For a given function $f : X \to \mathbb{R}$ interesting expectation values are for example mean and variance of the function f

$$\mu[f] = \mathrm{E}[f] = \int_X dx \, \pi(x) f(x), \tag{2.61a}$$

$$\sigma^2[f] = \mathrm{E}[(f - \mu[f])^2] = \int_X dx \, \pi(x)(f(x) - \mu[f])^2. \tag{2.61b}$$

If the dimension of X is large computation of these integrals by numerical quadrature (grid-based methods) is often infeasible. The number of operations (function evaluations, summations, etc.) grows exponentially with the dimension of the domain X (curse of dimensionality).

The central idea of Monte Carlo is to generate samples $x_1, \ldots, x_N$ distributed according to the target density π and to approximate the expectation values by the following unbiased sample estimators

$$m[f] = \frac{1}{N} \sum_{k=1}^{N} f(x_k), \tag{2.62a}$$

$$s^2[f] = \frac{1}{N-1} \sum_{k=1}^{N} (f(x_k) - m[f])^2. \tag{2.62b}$$

The error incurred by the approximation goes to zero in the large sample limit ($N \to \infty$).

Successive samples from a Markov chain are not independent so we can not compute the error of the mean using the central limit theorem (CLT). The CLT asserts that the sample mean $m[f]$ of N independent and identically distributed random variables with finite expectation $\mu[f]$ and variance $\sigma^2[f]$ follows a normal distribution around the expected value in the limit of large sample sizes ($N \to \infty$). The error of the mean is defined as the following standard deviation $\epsilon^2 = \mathrm{E}[(m[f] - \mu[f])^2]$ which asymptotically satisfies

$$\epsilon^2 = \frac{\sigma^2[f]}{N}. \tag{2.63}$$

A central limit theorem for reversible Markov chains [54] establishes an analogous result: The sample mean of a reversible Markov chain

is normally distributed around the true value asymptotically and the error of the sample mean ϵ is given by

$$\epsilon^2 = \frac{\sigma^2[f]}{N_{\text{eff}}}. \tag{2.64}$$

The effective sample size N_{eff} takes into account the dependency between samples from the Markov chain. It can be computed using the integrated autocorrelation time τ_{int} which reduces the actual sample size N according to

$$N_{\text{eff}} = \frac{N}{2\tau_{\text{int}}}. \tag{2.65}$$

The integrated autocorrelation time is defined as

$$\tau_{\text{int}} = \frac{1}{2} + \sum_{t=1}^{\infty} \Gamma[f](t). \tag{2.66}$$

The function $\Gamma[f](t)$ is the normalized autocorrelation function of the process $(f_t) = (f(X_t))$. For a stationary process it is defined as,

$$\Gamma[f](\tau) = \frac{\mathrm{E}[(f_t - \mathrm{E}[f])(f_{t+\tau} - \mathrm{E}[f])]}{\sigma^2[f]}. \tag{2.67}$$

In most cases the function $\Gamma[f]$ is approximated using a set of N samples of the Markov chain

$$a[f](\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} (f_t - m[f])(f_{t+\tau} - m[f]). \tag{2.68}$$

This estimate of $\Gamma[f]$ is not unbiased. The unbiased estimate is multiplied by an additional factor $1 - \tau/N$ which suppresses values at large values of τ. At large values of τ the estimate $a[f](\tau)$ will have large errors since only $(N - \tau)$ samples are used to compute the estimate.

Simple summation of $a[f](\tau)$ as in (2.66) is problematic since the variance of $a[f]$ does not go to zero as τ goes to infinity. A possible solution is to use an initial positive sequence estimator truncating the sum once $a[f](2m) + a[f](2m+1)$ becomes negative for the first time. If $\Gamma[f]$ can be described by a single exponential with decay time τ_{exp} then one can show that $\tau_{\text{int}} \approx \tau_{\text{exp}}$. An exponential fit to $a[f]$ can then be used to obtain τ_{int}. A more thorough discussion of sampling errors in Markov chain Monte Carlo simulations can be found in [37].

2.6   ENHANCED SAMPLING

The existence of metastable conformations makes it extremely difficult to obtain accurate estimates of their stationary probabilities or the transition probabilities between pairs of conformations from finite simulation data. Enhanced sampling techniques [42, 59, 104, 110, 111, 122]

can accelerate the computation of stationary probabilities significantly under certain circumstances.

If it is possible to find a reaction coordinate or collective variable, i.e. a function that parameterizes the transition between conformations then an efficient computation of the stationary probabilities can, in many cases, be accomplished via the umbrella sampling method [110].

Formally such a collective variable is given by a mapping from the high-dimensional space of molecular configurations X to the real numbers,

$$\Phi : X \to \mathbb{R}. \tag{2.69}$$

One can then define the equilibrium probability density of the collective variable

$$\pi(\xi) = \frac{1}{Z} \int_X dx \, \exp(-\beta V(x))\delta(\xi - \Phi(x)). \tag{2.70}$$

The precise meaning of the above integral is provided by the coarea formula [43]. The constant Z is the partition function ensuring correct normalization

$$Z = \int_X dx \, \exp(-\beta V(x)). \tag{2.71}$$

The free energy along the collective variable is then defined as

$$F(\xi) = -\frac{1}{\beta} \log\left(Z\pi(\xi)\right). \tag{2.72}$$

With the help of the free energy the probability density can be cast into the standard canonical form $\pi(\xi) = \exp(-\beta F(\xi))/Z$. For this reason $F(\xi)$ is often called potential of mean force [43]. If ξ is a meaningful reaction coordinate that parameterizes the transition between conformations then $F(\xi)$ has a barrier for every transition between conformations.

2.6.1   Umbrella sampling

A natural strategy for an enhanced sampling scheme is to force the system to sample the barrier regions which would otherwise be only very rarely visited. This can be achieved via an additional harmonic potential $u^{(\alpha)}$ that forces the system to sample the vicinity of a given value ξ_α of the reaction coordinate

$$u^{(\alpha)}(\xi) = \frac{k}{2}(\xi - \xi_\alpha)^2. \tag{2.73}$$

The stiffness of the forcing potential k has to be large enough to restrain the system in situations in which the free energy $F(\xi)$ has a steep slope in a neighborhood of the umbrella center ξ_α.

The probability density along ξ for the biased system $\pi^{(\alpha)}(\xi)$ can be related to the unbiased one via

$$\pi^{(\alpha)}(\xi) = \frac{Z}{Z^{(\alpha)}} \underbrace{\exp(-\beta u^{(\alpha)}(\xi))}_{U^{(\alpha)}(\xi)} \pi(\xi) \tag{2.74}$$

where $Z^{(\alpha)}$ is the partition function from (2.71) with $V(x)$ replaced by the biased potential $V^{(\alpha)}(x) = V(x) + u^{(\alpha)}(\Phi(x))$.

The biased density $\pi^{(\alpha)}$ can be estimated via umbrella sampling simulations. The finite sample estimate of $\pi^{(\alpha)}$ will only provide information about π in a neighborhood of ξ_α. We can recover π for a range of ξ values by joining information obtained from biased simulations with different centers $(\xi^{(\alpha)})_{\alpha=1}^M$.

Using (2.74) we can in principle compute the unbiased from the biased density. Unfortunately, we also need all fractions $Z/Z^{(\alpha)}$ which are almost never available *a priori*. Fortunately, we can still recover the unbiased density using a statistical reweighting method. But, sufficient overlap of the distributions sampled by adjacent umbrella simulations is necessary to obtain meaningful results from the reweighting method.

The umbrella sampling method can be extended to multidimensional collective variables $\Psi : X \to \mathbb{R}^n$, but the number of biased simulations required to cover the set of relevant values of the (multidimensional) collective variable can quickly become unmanageable - the sufficient overlap condition is also necessary in the multidimensional setting.

Finding an admissible value for the stiffness of the biasing potential can also be a problem since the maximum steepness of the free energy in the relevant range of values of the collective variable is not known *a priori*. Finally, finding a good collective variable can be very difficult - the mapping Φ has to encode a potentially complicated transition mechanism on a high-dimensional space of molecular configurations into a single degree of freedom.

2.6.2 *The weighted histogram analysis method*

In this section we discuss the weighted histogram analysis method (WHAM), one particular method for obtaining the unbiased density $\pi^{(0)}(\xi) = \pi(\xi)$ from a set of biased densities $(\pi^{(\alpha)}(\xi))_{\alpha=1}^M$. The biased densities are not directly available but have to be estimated from samples generated by enhanced sampling simulations.

In order to apply the weighted histogram analysis method (WHAM) [35, 58] we subdivide the relevant interval of values of the collective variable Φ into a number of bins $[\xi_0, \xi_1), [\xi_1, \xi_2), \ldots, [\xi_{d-1}, \xi_d)$. Then we approximate all densities by densities that are piecewise constant over the bin, i.e. $\pi^{(\alpha)}(\xi) \approx \pi_i^{(\alpha)}(\xi))$ for all $\xi \in [\xi_{i-1}, \xi_i)$. Each density $\pi^\alpha(\xi)$ is thus approximated by a probability vector $(\pi_i^{(\alpha)})$. The biasing factors have also to be approximated by a piecewise constant function

with values given by the vector $(U_i^{(\alpha)})$. We denote by $c_i^{(\alpha)}$ the number of samples from a simulation at condition α belonging to the i-th bin $[\xi_{i-1}, \xi_i)$.

The likelihood of observing the counts $\{(c_i^{(1)})_{i=1}^d, \ldots, (c_i^{(M)})_{i=1}^d\}$ from independent simulations generating independent samples from the biased densities is then given by

$$
\begin{aligned}
\mathbb{P}(c^{(1)}, \ldots, c^{(M)} | \pi^{(1)}, \ldots, \pi^{(M)}) &= \prod_{\alpha=1}^{M} \mathbb{P}(c^{(\alpha)} | \pi^{(\alpha)}) \\
&= \prod_{\alpha=1}^{M} \prod_{i=1}^{d} (\pi_i^{(\alpha)})^{c_i^{(\alpha)}}.
\end{aligned}
\tag{2.75}
$$

The maximum likelihood estimation problem for WHAM is then

$$
\begin{aligned}
\max_{w^{(0)}, X^{(\alpha)}} \quad &\sum_{\alpha} \left(\sum_j c_j^{(\alpha)} \log \pi_j^{(\alpha)} \right) \\
\text{subject to} \quad &\sum_j \pi_j^{(\alpha)} = 1, \quad \pi_j^{(\alpha)} = X^{(\alpha)} U_j^{(\alpha)} w_j^{(0)}, \\
&X^{(\alpha)} > 0, \quad w_j^{(0)} > 0.
\end{aligned}
\tag{2.76}
$$

Where $X^{(\alpha)} = (Z^{(\alpha)})^{-1}$ and $w_j^{(0)} \propto Z \pi_j^{(0)}$ are the not necessarily normalized weights for the unbiased ensemble.

Ignoring all inequality constraints and substituting $\pi^{(\alpha)}$ via the reweighting condition we have the following Lagrangian for the equality constrained problem,

$$
\begin{aligned}
\mathcal{L} =& \sum_{\alpha} \sum_j c_j^{(\alpha)} \log \left(X^{(\alpha)} U_j^{(\alpha)} w_j^{(0)} \right) \\
&+ \sum_{\alpha} \lambda^{(\alpha)} \left(\sum_j X^{(\alpha)} U_j^{(\alpha)} w_j^{(0)} - 1 \right).
\end{aligned}
\tag{2.77}
$$

Necessary conditions for a local optimum of (2.76) are

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial X^{(\alpha)}} &= \frac{\sum_j c_j^{(\alpha)}}{X^{(\alpha)}} + \lambda^{(\alpha)} \sum_j U_j^{(\alpha)} w_j^{(0)} = 0, \\
\frac{\partial \mathcal{L}}{\partial w_k^{(0)}} &= \frac{\sum_{\alpha} c_k^{(\alpha)}}{w_k^{(0)}} + \sum_{\alpha} \lambda^{(\alpha)} X^{(\alpha)} U_k^{(\alpha)} = 0.
\end{aligned}
\tag{2.78}
$$

Rearranging expressions and using the normalization condition we have

$$
\begin{aligned}
\lambda^{(\alpha)} &= -\sum_j c_j^{(\alpha)}, \\
w_k^{(0)} &= \frac{\sum_{\alpha} c_k^{(\alpha)}}{\sum_{\alpha} \left(\sum_j c_j^{(\alpha)} \right) X^{(\alpha)} U_k^{(\alpha)}}.
\end{aligned}
\tag{2.79}
$$

Plugging the expression for λ_α into the first equation of (2.78) and using it together with the second equation of (2.79) yields the set of WHAM-equations, which are usually solved via a self-consistent iteration,

$$X^{(\alpha)} = \frac{1}{\sum_j U_j^{(\alpha)} w_j^{(0)}},$$

$$w_k^{(0)} = \frac{\sum_\alpha c_k^{(\alpha)}}{\sum_\alpha \left(\sum_j c_j^{(\alpha)}\right) X^{(\alpha)} U_k^{(\alpha)}}. \tag{2.80}$$

Equation (2.76) is invariant under the change of variables $\tilde{w}_j^{(0)} = A w_j^{(0)}$ and $\tilde{X}^{(\alpha)} = A^{-1} X^{(\alpha)}$ with $A > 0$, so that the solution to (2.76) is not unique. This is also true for the WHAM-equations (2.80). But, the self-consistent iteration has to be started from an initial guess which removes this ambiguity. Still, care has to be taken to avoid over- or underflow if a self-consistent iteration for (2.80) is implemented. It is an open question how the choice of scaling affects the performance of the algorithm and how to choose an optimal scaling. Below we show how the self-consistent iteration of (2.80) can be avoided when solving the WHAM problem (2.76) using a method outlined in [127].

The second equation in (2.79) can be used to eliminate the variables $w_k^{(0)}$. This results in the following minimization problem

$$\min_{X^{(\alpha)}} \quad -\sum_\alpha \sum_j c_j^{(\alpha)} \left(\log X^{(\alpha)} - \log\left(\sum_\beta \sum_k c_k^{(\beta)} X^{(\beta)} U_j^{(\beta)}\right)\right) \tag{2.81}$$

$$\text{s. t.} \quad X^{(\alpha)} > 0.$$

It can be seen that (2.81) is still invariant under a global scaling, $\tilde{X}^{(\alpha)} = A^{-1} X^{(\alpha)}$. This invariance can be removed by requiring, for example, $X^{(1)} = 1$.

Introducing the counts per state $n_j = \sum_\alpha c_j^{(\alpha)}$ and the counts per ensemble $m_\alpha = \sum_j c_j^{(\alpha)}$, the piecewise constant biasing potentials $u_{\alpha,k} = \log U_k^{(\alpha)}$, and the log-space variables $x_\alpha = \log X^{(\alpha)}$ we can transform (2.81) to

$$\min_x \quad -\sum_\alpha m_\alpha x_\alpha + \sum_j n_j \log\left(\sum_\beta \exp(x_\beta + u_{\beta,j} + \log m_\beta)\right) \tag{2.82}$$

$$\text{s. t.} \quad x_1 = 0$$

where we have added the constraint $x_1 = 0$ to enforce a unique solution. The problem (2.82) is a constrained convex optimization problem which admits a unique solution. Once the optimal solution, x_α^* to (2.82) has been found the desired unbiased weights $w_k^{(0)}$ can be computed according to the second equation of (2.80),

$$w_k^{(0)} = \frac{n_k}{\sum_\alpha \exp\left(x_\alpha + u_{\alpha,k} + \log m_\alpha\right)}. \tag{2.83}$$

A remarkable feature of the minimization problem in (2.82) is that the number of unknowns is completely independent of the number of bins that were used for the discretization of the collective variable. But, the number of samples per umbrella necessary to visit all relevant bins with the right frequency will still be sensitive to the total number of bins.

The method can be easily extended to cover also a multidimensional collective variable [57], but as discussed above the computational cost for an exhaustive sampling of the relevant set of values can quickly become a computational bottleneck.

The WHAM approach is not limited to umbrella sampling data. Any set of enhanced sampling simulations can be analyzed as long as a meaningful positive value for the reweighting transformation $U_i^{(\alpha)}$ can be determined for all bins. Obviously, this excludes situations in which $U^{(\alpha)}(\xi)$ is not approximately constant over the chosen bins. Reweighting methods that do not require binning of the data have also been developed to circumvent this problem [99, 106, 107, 121].

A severe limitation of the method is that samples have to be independent. In most cases samples will be generated by MCMC or MD simulations which generated correlated samples. As a result simulations have to be long enough to serve as a source of approximately independent samples which can prevent the use of many short trajectories which can be efficiently simulated in parallel.

The discrete transition matrix reweighting method (dTRAM), a state discrete reweighting method that can handle correlated samples is described in [123, 124]. The method uses reversible Markov state models for the reweighting procedure. The method is outlined below and an algorithm for the efficient solution of the dTRAM problem is derived.

ESTIMATION OF REVERSIBLE MARKOV STATE MODELS BY CONVEX-CONCAVE PROGRAMS

In this chapter efficient algorithms for the estimation of reversible transition matrices given observations of a Markov chain are developed.

Maximum likelihood estimation of transition probabilities satisfying a detailed balance equation leads to a nonlinear program which is difficult to solve. The problem is nonconvex and the number of unknowns in the problem scales quadratically in the number of states of the Markov chain. Previously, the problem was solved by a self-consistent iteration [12, 84, 123] which can take a long time to converge and whose convergence behavior can be sensitive to the data that is used.

It is shown that a duality argument from [123] can be used to transform the maximum likelihood estimation problem into a convex-concave program which can be solved using a primal-dual interior-point method for variational inequalities [86]. The resulting algorithm is superlinearly convergent, robust with respect to the input data, and can be applied to Markov chains with a large number of states. The algorithmic approach allows to tackle a number of problems related to the estimation of reversible transition matrices, e.g. situations in which *a priori* information about stationary probabilities of the chain is available.

Reversible transition probabilities for an unbiased ensemble can be estimated from a combination of enhanced sampling simulations and standard (unbiased) molecular dynamics simulations using the discrete transition matrix reweighting analysis method (dTRAM) [123, 124].

The proposed method for the estimation of reversible transition probabilities is extended to also cover the dTRAM problem. The algorithm can efficiently compute the dTRAM estimator for enhanced sampling simulations at many different biasing conditions and for Markov chains on large state spaces. It is shown that the developed algorithm can solve the dTRAM problem orders of magnitudes faster than the self-consistent iteration proposed in [123]. The material presented in this chapter has previously appeared in [114].

3.1 MARKOV CHAIN ESTIMATION

A Markov chain on a finite state space is completely characterized by a square matrix of conditional probabilities, $P = (p_{ij}) \in \mathbb{R}^{n \times n}$. The entry p_{ij} is the probability for the chain to make a transition to state j given that it currently resides in state i. The matrix P is stochastic, i.e. $\sum_j p_{ij} = 1$ for all i. If P is irreducible then there exists a unique vector, $\pi = (\pi_i) \in \mathbb{R}^n$, of positive probabilities such that π is invariant

under the action of P, $\pi^T P = \pi^T$. The vector π is called the stationary vector of the chain.

If there is a vector, π, of probabilities for which P fulfills the following detailed balance condition,

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{3.1}$$

then the chain is a reversible Markov chain with stationary vector π, [61].

In Markov chain estimation one is interested in finding an optimal transition matrix estimate P from a given finite observation $X = \{X_0, X_1, \ldots, X_N\}$ of a Markov chain with unknown transition matrix. The matrix of transition counts $C = (c_{ij})$ together with the initial state $X_0 = x_0$ is a minimal sufficient statistics for the transition matrix [2, 22]. The element c_{ij} denotes the observed number of transitions between state i and state j in X. The matrix P is optimal if it maximizes the following log-likelihood

$$L(C|P) = \sum_{i,j} c_{ij} \log p_{ij}. \tag{3.2}$$

For finite ensembles consisting of finite length observations one can simply add the matrices of transition counts for each observation. The accumulated counts together with the empirical measure of the initial states is then a sufficient statistics for the finite ensemble of observations.

For reversible Markov chain estimation one constrains the general Markov chain maximum likelihood estimation (MLE) problem to the set of all stochastic matrices for which detailed balance with respect to some vector of probabilities holds. Thus we can find the reversible MLE transition matrix from the following nonlinear program,

$$\min_{\pi, P} \quad -\sum_{i,j} c_{ij} \log p_{ij}$$

$$\text{subject to} \quad p_{ij} \geq 0, \ \sum_j p_{ij} = 1, \ \pi_i > 0, \ \sum_i \pi_i = 1, \tag{3.3}$$

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

In [123, 124] problem (3.3) has been extended to the discrete transition matrix reweighting analysis method (dTRAM). For dTRAM, simulation data at multiple thermodynamic states $\alpha = 0, \ldots, M$ is collected in order to efficiently estimate the stationary vector at the unbiased condition, $\alpha = 0$. A positive reweighting transformation relates the stationary vector at the biased condition, $\alpha > 0$, to the stationary vector at the unbiased condition,

$$\pi_i^{(\alpha)} = U_i^{(\alpha)} \pi_i^{(0)} = \exp(u_i^{(\alpha)}) \pi_i^{(0)}. \tag{3.4}$$

This coupling allows us to combine the information from all ensembles into the estimate for $\pi^{(0)}$.

The dTRAM problem consists of reversible MLE problems for each thermodynamic state coupled via the reweighting transformation (3.4). The desired stationary vector can be obtained as the optimal point of the following nonlinear program,

$$\min_{\pi^{(\alpha)}, P^{(\alpha)}} \quad -\sum_\alpha \sum_{i,j} c_{ij}^{(\alpha)} \log p_{ij}^{(\alpha)}$$

$$\text{subject to} \quad p_{ij}^{(\alpha)} \geq 0, \ \sum_j p_{ij}^{(\alpha)} = 1, \ \pi_i^{(\alpha)} > 0, \ \sum_i \pi_i^{(\alpha)} = 1, \quad (3.5)$$

$$\pi_i^{(\alpha)} p_{ij}^{(\alpha)} = \pi_j^{(\alpha)} p_{ji}^{(\alpha)}, \ \pi_i^{(\alpha)} = U_i^{(\alpha)} \pi_i^{(0)}.$$

We show that the convex-concave reformulation of the reversible MLE problem can be extended to derive an efficient numerical algorithm for the solution of the dTRAM problem. Additional structure in the linear systems arising during the primal-dual iteration can be used so that the problem can be solved efficiently for many coupled chains.

3.2   DUAL OF THE REVERSIBLE MLE PROBLEM

In [123] a duality argument was used to show that finding the MLE of (3.3) for given positive weights π_i is equivalent to the following concave maximization problem,

$$\max_x \quad \sum_{i,j} c_{ij} \log(\pi_i x_j + \pi_j x_i) - \sum_{i,j} c_{ij} \log \pi_j - \sum_i x_i \qquad (3.6)$$

$$\text{subject to} \quad x_i \geq 0.$$

The x_i correspond to the Lagrange multipliers for the row normalization constraint in the primal problem (3.3). The optimal transition probabilities can be recovered according to

$$p_{ij}^* = \frac{(c_{ij} + c_{ji})\pi_j}{\pi_i x_j^* + \pi_j x_i^*}, \quad j \neq i. \qquad (3.7)$$

The vector x^* denotes the optimal point of (3.6) and the diagonal entries p_{ii}^* are determined by the row normalization condition. It is clear that p_{ij}^* is a proper probability irrespective of the normalization of the weights since any scaling of π_i cancels out in (3.7).

The problem (3.3) with fixed stationary vector is a convex minimization problem. To show that strong duality holds we need to verify Slater's condition [13]. Since all inequality constraints are linear it suffices to find a point p_{ij}^* for which the objective function in (3.3) is bounded and

$$p_{ij}^* \geq 0, \ \sum_j p_{ij}^* = 1, \ \pi_i p_{ij}^* = \pi_j p_{ji}^*. \qquad (3.8)$$

Such a point can be constructed using the Metropolis-Hastings acceptance criterion. Let

$$q_{ij} = \frac{c_{ij}}{\sum_k c_{ik}}, \ \alpha_{ij} = \min\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\}, \ p_{ij}^* = \alpha_{ij} q_{ij}, \ j \neq i, \qquad (3.9)$$

and $p_{ii}^* = 1 - \sum_{j \neq i} p_{ij}^*$. Then (p_{ij}^*) satisfies the conditions in (3.8) and the value of the objective function in (3.3) is finite. Thus we can obtain the optimal value of the primal problem from the dual function. As usual we compute the dual function

$$g_\pi(\lambda, x) = \min_P L_\pi(P, \lambda, x) \tag{3.10}$$

using the Lagrange function of the primal problem,

$$\begin{aligned} L_\pi(P, \lambda, x) &= -\sum_{ij} c_{ij} \log p_{ij} + \sum_{ij} \lambda_{ij} (\pi_i p_{ij} - \pi_j p_{ji}) \\ &\quad + \sum_{ij} x_i (p_{ij} - 1) \\ &= -\sum_{ij} c_{ij} \log p_{ij} + \sum_{ij} (\pi_i (\lambda_{ij} - \lambda_{ji}) + x_i) p_{ij} \\ &\quad - \sum_i x_i. \end{aligned} \tag{3.11}$$

The dual function is bounded from below only if $x_i \geq 0$.

In [123] the inequality constraints on x_i were not made explicit. The non-negativity requirement can be seen from the following splitting of the Lagrangian,

$$\begin{aligned} L_\pi(P, \lambda, x) &= -\sum_{i,j \in I} c_{ij} \log p_{ij} + \sum_{i,j \in I} (\pi_i (\lambda_{ij} - \lambda_{ji}) + x_i) p_{ij} \\ &\quad + \sum_{i,j \notin I} (\pi_i (\lambda_{ij} - \lambda_{ji}) + x_i) p_{ij} - \sum_i x_i \end{aligned} \tag{3.12}$$

with index set $I = \{(i,j) | c_{ij} > 0\}$ and the constraint $p_{ij} \geq 0$. The value $\min_P L_\pi$ is not bounded from below if $\pi_i(\lambda_{ij} - \lambda_{ji}) + x_i < 0$ for some $(i,j) \notin I$. Therefore $x_i \geq 0$ for all $(i,i) \notin I$. It is also not bounded from below if $\pi_i(\lambda_{ij} - \lambda_{ji}) + x_i \leq 0$ for some $(i,j) \in I$, so that $x_i > 0$ for all $(i,i) \in I$.

The first order condition for $\min_P L_\pi$ is

$$\frac{\partial L_\pi}{\partial p_{ij}} = -\frac{c_{ij}}{p_{ij}} + \pi_i(\lambda_{ij} - \lambda_{ji}) + x_i = 0. \tag{3.13}$$

The optimal point p_{ij} is thus given in terms of the multipliers λ and x,

$$p_{ij} = \frac{c_{ij}}{\pi_i(\lambda_{ij} - \lambda_{ji}) + x_i}. \tag{3.14}$$

Using (3.14) we obtain an analytic expression for the dual function

$$g_\pi(\lambda, x) = \sum_{ij} c_{ij} \log(\pi_i(\lambda_{ij} - \lambda_{ji}) + x_i) - \sum_i x_i + \text{const.} \tag{3.15}$$

The first order condition for $\max_\lambda g_\pi$ is

$$\frac{\partial g}{\partial \lambda_{ij}} = \frac{\pi_i c_{ij}}{\pi_i(\lambda_{ij} - \lambda_{ji}) + x_i} - \frac{\pi_j c_{ji}}{\pi_j(\lambda_{ji} - \lambda_{ij}) + x_j} = 0 \tag{3.16}$$

and the optimal λ satisfies

$$\lambda_{ij} - \lambda_{ji} = \frac{\pi_i c_{ij} x_j - \pi_j c_{ji} x_i}{\pi_i \pi_j (c_{ij} + c_{ji})}. \tag{3.17}$$

Plugging (3.17) into (3.14) we obtain the expression for the optimal transition probabilities in (3.7) and we can express the optimal dual value $\max_\lambda g(\lambda, x)$ as

$$f(\pi, x) = \sum_{ij} \log(\pi_i x_j + \pi_j x_i) - \sum_{ij} c_{ij} \log \pi_j - \sum_i x_i. \tag{3.18}$$

Using the function f in (3.18) the reformulation of the reversible MLE problem, (3.3), as a saddle-point problem with constraints is

$$\min_\pi \max_x \quad \sum_{i,j} c_{ij} \log(\pi_i x_j + \pi_j x_i) - \sum_{i,j} c_{ij} \log \pi_j - \sum_i x_i$$
$$\text{subject to} \quad x_i \geq 0, \ \pi_i > 0, \ \sum_i \pi_i = 1. \tag{3.19}$$

The objective in (3.19) is concave in x but non-convex in π. The problem can however be easily cast into a convex-concave form by the following change of variables,

$$\pi_i \propto e^{y_i}, \tag{3.20}$$

and by replacing the normalization condition with the simpler constraint

$$y_1 = 0. \tag{3.21}$$

The constraint in (3.21) removes the invariance of the objective in (3.19) with respect to a constant shift of y. Proper stationary probabilities π_i can be obtained from the new variables y_i according to (3.20) followed by straightforward normalization. The variable y_i is the negative free energy of the state i.

The final form of the dual reversible MLE problem is

$$\max_y \min_x \quad -\sum_{i,j} c_{ij} \log \left(x_i e^{y_j} + x_j e^{y_i} \right) + \sum_i x_i + \sum_{i,j} c_{ij} y_j$$
$$\text{subject to} \quad x_i \geq 0, \ y_1 = 0. \tag{3.22}$$

The objective in (3.22) is convex in x and concave in y. The feasible set is convex so that (3.22) is a convex-concave program.

For a given state space with n states the original reversible MLE problem (3.3), a non-convex constrained minimization problem in $\mathcal{O}(n^2)$ unknowns, is reduced to a convex-concave programming problem in $\mathcal{O}(n)$ unknowns with simple constraints.

3.2.1 *Scaling*

We observe that the number of iterations needed for the solution of (3.22) using the algorithm from [86] can be drastically reduced by scaling the count-matrix by a constant factor γ chosen as

$$\gamma = \left(\max_{i,j} c_{ij} \right)^{-1}. \tag{3.23}$$

With scaled entries $\tilde{c}_{ij} = \gamma c_{ij}$ and scaled variables $\tilde{x} = \gamma x$, $\tilde{y} = y$ we have

$$\tilde{f}(\tilde{x}, \tilde{y}) = \gamma f(x, y) + \text{const.} \tag{3.24}$$

The constraints in (3.22) are invariant under the scaling so that the optimal point for (3.22) can be obtained from the optimal solution to the scaled problem.

The resulting stationary probabilities as well as the transition probabilities are invariant under the scaling,

$$\tilde{p}_{ij} = \frac{(\tilde{c}_{ij} + \tilde{c}_{ji})e^{\tilde{y}_j}}{\tilde{x}_i e^{\tilde{y}_j} + \tilde{x}_j e^{\tilde{y}_i}} = \frac{(c_{ij} + c_{ji})e^{y_j}}{x_i e^{y_j} + x_j e^{y_i}} = p_{ij}. \tag{3.25}$$

3.2.2 *Special cases and extensions*

The reversible estimation problem with fixed stationary vector π

$$\begin{aligned}
\min_{P} \quad & -\sum_{i,j} c_{ij} \log p_{ij} \\
\text{subject to} \quad & p_{ij} \geq 0, \ \sum_j p_{ij} = 1, \ \pi_i p_{ij} = \pi_j p_{ji}
\end{aligned} \tag{3.26}$$

is a convex problem and can efficiently be solved in its dual formulation (3.6) using an interior-point method for convex programming problems.

The reversible estimation problem with partial information about the stationary vector

$$\begin{aligned}
\min_{\pi, P} \quad & -\sum_{i,j} c_{ij} \log p_{ij} \\
\text{subject to} \quad & p_{ij} \geq 0, \ \sum_j p_{ij} = 1, \ \pi_i > 0, \ \sum_i \pi_i = 1, \\
& \pi_i p_{ij} = \pi_j p_{ji}, \ \pi_i = \nu_i \ i \in I,
\end{aligned} \tag{3.27}$$

with I not $\subseteq \{1, \ldots, n\}$ and given positive weights $(\nu_i)_{i \in I}$ can be solved via its dual

$$\begin{aligned}
\max_{y} \min_{x} \quad & -\sum_{i,j} c_{ij} \log \left(x_i e^{y_j} + x_j e^{y_i} \right) + \sum_i x_i + \sum_{i,j} c_{ij} y_j \\
\text{subject to} \quad & x_i \geq 0, \ y_i = \log \nu_i \ i \in I.
\end{aligned} \tag{3.28}$$

The reversible estimation problem with bound-constrained information about the stationary vector

$$
\min_{\pi,P} \quad -\sum_{i,j} c_{ij} \log p_{ij}
$$

$$
\text{subject to} \quad p_{ij} \geq 0, \ \sum_j p_{ij} = 1, \ \pi_i > 0, \ \sum_i \pi_i = 1, \qquad (3.29)
$$

$$
\pi_i p_{ij} = \pi_j p_{ji}, \ \eta_i \leq \pi_i \leq \xi_i \ i \in I.
$$

with $I \subseteq \{1, \ldots, n\}$ and given positive bounds $(\eta_i)_{i \in I}$, $(\xi_i)_{i \in I}$ can be solved via the dual

$$
\max_y \min_x \quad -\sum_{i,j} c_{ij} \log \left(x_i e^{y_j} + x_j e^{y_i} \right) + \sum_i x_i + \sum_{i,j} c_{ij} y_j
$$

$$
\text{subject to} \quad x_i \geq 0, \ \log \eta_i \leq y_i \leq \log \xi_i \ i \in I. \tag{3.30}
$$

The two problems (3.28), (3.30) are convex-concave programming problems. Nonlinear, convex inequality and linear equality constraints possibly coupling x and y can also be treated within the algorithmic framework of [86]. A special case with possible interest for applications are bound constraints on the integrated stationary weights on subsets $S \subseteq \{1, \ldots, n\}$,

$$
\sum_{i \in S} \pi_i \leq \nu. \tag{3.31}
$$

Equation (3.31) can be expressed in terms of variables y_i as

$$
\log \sum_{i \in S} e^{y_i} \leq \log \nu_k, \tag{3.32}
$$

The logarithm of a sum of exponentials is a convex function, [13].

3.2.3 dTRAM

We can apply the duality argument to each thermodynamic state in (3.5) and introduce the coupling between different ensembles, (3.4), through linear equality constraints. The resulting convex-concave programming problem is

$$
\max_{y^{(\alpha)}} \min_{x^{(\alpha)}} \quad -\sum_{\alpha} \sum_{i,j} c_{ij} \log \left(x_i^{(\alpha)} e^{y_j^{(\alpha)}} + x_j^{(\alpha)} e^{y_i^{(\alpha)}} \right) + \sum_i x_i^{(\alpha)} + \sum_{i,j} c_{ij} y_j^{(\alpha)}
$$

$$
\text{subject to} \quad x_i^{(\alpha)} \geq 0, \ y_i^{(\alpha)} - y_i^{(0)} = u_i^{(\alpha)}, \ y_1^{(0)} = 0.
$$

$$
\tag{3.33}
$$

The number of iterations required to solve the dTRAM problem is also greatly reduced by scaling each count-matrix according to

$$
\tilde{c}_{ij}^{(\alpha)} = \gamma c_{ij}^{(\alpha)} \tag{3.34}
$$

with

$$\gamma = \max_{\alpha,i,j} c_{ij}^{(\alpha)} \tag{3.35}$$

As for the reversible MLE problem a larger class of related dTRAM problems can be solved by augmenting the dual problem (3.33) with convex constraints, e.g. dTRAM with partial or bound constrained information about the unbiased stationary vector. It must be ensured that the additional constraints on the biased stationary probabilities do not result in an infeasible problem, i.e. the reweighting condition (3.4) and the constraints cannot be fulfilled simultaneously.

3.3   CONVEX-CONCAVE PROGRAMS

A convex-concave program is the following saddle point problem,

$$\begin{aligned}
&\max_{y} \min_{x} \quad f(x,y) \\
&\text{subject to} \quad (x,y) \in \mathcal{K}
\end{aligned} \tag{3.36}$$

with f convex in x, concave in y, and $\mathcal{K} \subseteq \mathbb{R}^n$ a convex set.

Convex-concave programs can be treated as special cases of finite-dimensional variational inequality (VI) problems, [32]: For a given feasible set $\mathcal{K} \subseteq \mathbb{R}^n$ and a mapping $\Phi : \mathcal{K} \to \mathbb{R}^n$ find a point $z^* \in \mathcal{K}$ such that

$$(z - z^*)^T \Phi(z^*) \geq 0 \quad \forall z \in \mathcal{K}. \tag{3.37}$$

Any point z^* satisfying (3.37) is a solution or optimal point for the VI. The convex-concave program is cast into the VI-form by defining

$$\Phi(z) = \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}, \quad z = (x,y). \tag{3.38}$$

A mapping Φ is said to be monotone if

$$(z' - z)^T (\Phi(z') - \Phi(z)) \geq 0 \quad \forall z', z \in \mathcal{K}. \tag{3.39}$$

Monotonicity of (3.38) follows from the convex-concave property of f.

If $\mathcal{K}$ is a convex polyhedral set, i.e. solely defined in terms of linear equalities and inequalities,

$$\mathcal{K} = \{z \in \mathbb{R}^n | Az - b = 0, \ Gz - h \leq 0\}, \tag{3.40}$$

then z solves the VI (3.37) if and only if there are vectors λ, ν, s, such that the following Karush-Kuhn-Tucker (KKT) conditions are fulfilled [32],

$$\begin{aligned}
\Phi(z) + A^T \nu + G^T \lambda &= 0 \\
Az - b &= 0 \\
Gz - h + s &= 0 \\
\lambda^T s &= 0 \\
\lambda, s &\geq 0
\end{aligned} \tag{3.41}$$

The vectors λ and ν are dual variables associated with the inequality and equality constraints. The vector of slack variables, $s = (h - Gz)$, transforms the linear inequality constraints for z into simple non-negativity constraints for s. Optimality conditions for convex $\mathcal{K}$ in standard form, i.e. defined by a finite number of linear equalities and convex inequalities, are also available, cf. [32].

A direct application of a Newton type method to (3.41) ensuring positivity of λ and s is usually unsuccessful since the solution progress rapidly stagnates once the iterates approach the boundary of the feasible set.

A possible strategy to circumvent this problem is numerical path-following. Instead of attempting a direct solution of (3.41) path-following proceeds by solving a sequence of problems with perturbed complementarity condition,

$$
\begin{aligned}
\Phi(z) + A^T\nu + G^T\lambda &= 0 \\
Az - b &= 0 \\
Gz - h + s &= 0 \\
\lambda^T s &= \mu \\
\lambda, s &\geq 0
\end{aligned}
\tag{3.42}
$$

tracing the central path of solutions $z^*(\mu)$ towards $z^*(0)$ with $\mu \to 0^+$. Perturbing the complementarity condition ensures that the boundary of the feasible set is not reached prematurely and the iteration makes good progress along the computed search direction.

Interior-point methods ensure the positivity of λ and s at each step of the iteration. If in addition a strictly feasible starting point $Az^{(0)} - b = 0$, $Gz^{(0)} - h + s^{(0)} = 0$ is used then all iterates produced by the algorithm lie in the interior of the feasible region.

Progress towards a solution of the perturbed KKT conditions (3.42) is usually made by taking steps along the Newton direction computed from the following linear system,

$$
\begin{pmatrix} D\Phi(z) & A^T & G^T & 0 \\ A & 0 & 0 & 0 \\ G & 0 & 0 & I \\ 0 & 0 & S & \Lambda \end{pmatrix} \begin{pmatrix} \Delta z \\ \Delta \nu \\ \Delta \lambda \\ \Delta s \end{pmatrix} = - \begin{pmatrix} \Phi(z) + A^T\nu + G^T\lambda \\ Az - b \\ Gz - h + s \\ S\Lambda \mathbf{e} - \mu\mathbf{e} \end{pmatrix},
\tag{3.43}
$$

with diagonal matrices $S = \mathrm{diag}(s_1, s_2, \dots)$, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots)$, the vector $\mathbf{e} = (1, 1, \dots)$, and the perturbation parameter $\mu > 0$.

We use the following short-hand notation for the dual residuum,

$$
r_d = \Phi(z) + A^T\nu + G^T\lambda,
\tag{3.44}
$$

the primal residuals,

$$
\begin{aligned}
r_{p,1} &= Az - b, \\
r_{p,2} &= Gz - h + s,
\end{aligned}
\tag{3.45}
$$

and the perturbed complementary slackness,

$$r_c(\mu) = S\Lambda\mathbf{e} - \mu\mathbf{e}. \tag{3.46}$$

Solving the linear system (3.43) is the most expensive part of the algorithm. The sparse block structure of (3.43) can be used to significantly speed up the solution process. Elimination of Δs and $\Delta\lambda$ reduces (3.43) to the *augmented system*

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta z \\ \Delta\nu \end{pmatrix} = -\begin{pmatrix} r_d + G^T\Sigma r_{p,2} - G^T S^{-1} r_c(\mu) \\ r_{p,1} \end{pmatrix}, \tag{3.47}$$

with diagonal matrix $\Sigma = S^{-1}\Lambda$ and augmented Jacobian $H = D\Phi + G^T\Sigma G$. The increments $\Delta\lambda$ and Δs can be computed from Δz,

$$\begin{aligned} \Delta s &= -r_{p,2} - G\Delta z \\ \Delta\lambda &= -\Sigma\Delta s - S^{-1} r_c(\mu). \end{aligned} \tag{3.48}$$

For nonsingular H further elimination of Δz from (3.47) is possible. The resulting *normal equations* for $\Delta\nu$ are,

$$S\Delta\nu = r_2 - AH^{-1}r_1. \tag{3.49}$$

The vectors r_i are the two components of the RHS of (3.47) and the matrix $S = \left(AH^{-1}A^T\right)$ is the Schur complement of H. The increment Δz can then be computed according to

$$\Delta z = -H^{-1}(r_1 + A^T\Delta\nu). \tag{3.50}$$

A singular matrix H can for example occur for an equality-constrained convex programming problem for which the objective is not strictly convex. Even if the constraints ensure that the problem has a unique solution, H will be singular so that the normal equations can not be formed.

For convex programming problems a non-singular H can be efficiently factorized using a symmetric positive-definite Cholesky factorization. In the convex-concave case the Jacobian of the mapping Φ is not symmetric,

$$D\Phi(z) = \begin{pmatrix} \nabla_x\nabla_x f(x,y) & \nabla_y\nabla_x f(x,y)^T \\ -\nabla_y\nabla_x f(x,y) & -\nabla_y\nabla_y f(x,y) \end{pmatrix}. \tag{3.51}$$

In that case the augmented system is not symmetric and the Cholesky factorization can not be used.

A further speed-up in the computation of the Newton direction can be achieved through the exploitation of sparse or block-sparse structure possibly present in $D\Phi$, G, A. In this situation solution via an iterative method can be particularly efficient if a good preconditioner is available.

3.4   THE RALPH-WRIGHT ALGORITHM

A primal-dual interior-point algorithm for the solution of monotone variational inequalities is developed in [86]. The algorithm uses the Newton system (3.43) to compute the search direction. At each step of the iteration the following complementarity measure,

$$\tilde{\mu} = \frac{\lambda^T s}{m} \tag{3.52}$$

is computed from the current iterate (z, λ, ν, s). The integer m denotes the total number of inequality constraints defining $\mathcal{K}$. The scalar $\tilde{\mu}$ measures the average violation of the complementarity condition $\lambda^T s = 0$ in (3.41). The perturbation parameter for the computation of the search direction is then set to

$$\mu = \sigma \tilde{\mu} \tag{3.53}$$

with $\sigma \in [0, \frac{1}{2}]$.

At each step of the iteration the affine scaling direction, $\sigma = 0$, is computed. A line-search ensures that the step remains in a neighborhood $\mathcal{N}$ of the central path. The step is accepted only if it produces a sufficiently large reduction of $\tilde{\mu}$.

If the affine scaling step is unsuccessful the partial centering direction, $\sigma \in (0, \frac{1}{2}]$, is computed. The line-search ensures an Armijo-type sufficient decrease condition for $\tilde{\mu}$ while confining the new iterate to $\mathcal{N}$.

The partial-centering step, aims at an equlibration of all complementarity products to a fraction of the current average complementarity, $\tilde{\mu}$, while the affine-scaling step tries to reduce all complementarity products to zero. Both steps aim to reduce the primal and dual residuals to zero.

The formal definition of the neighborhood $\mathcal{N}$ can be found in [86]. In addition to the positivity of λ, s it is essentially characterized by the requirement that the norm of the primal and dual residuals are bounded from above by a multiple of $\tilde{\mu}$ and that the minimum over all complementarity products, $\lambda_i s_i$, is bounded from below by a fraction of the complementarity measure $\tilde{\mu}$.

The matrix on left-hand side of (3.43) is independent of μ. Solutions for different values of μ are cheaply available once the matrix on the left-hand side has been factorized.

3.5   IMPLEMENTATION DETAILS

In order to apply the algorithm in [86] to the reversible MLE problem (3.22) we transform the convex-concave program into the VI form using

the mapping $\Phi = (\nabla_x f, -\nabla_y f)$ in (3.38). The gradient of the objective in (3.22) is given by

$$
\begin{aligned}
\partial_{x_k} f &= -\sum_j \frac{(c_{kj} + c_{jk})e^{y_j}}{x_k e^{y_j} + x_j e^{y_k}} + 1 \\
\partial_{y_k} f &= -\sum_j \frac{(c_{kj} + c_{jk})x_j e^{y_k}}{x_k e^{y_j} + x_j e^{y_k}} + \sum_i c_{ik}.
\end{aligned}
\tag{3.54}
$$

For the computation of the Newton direction we also need the Jacobian $D\Phi$. The diagonal blocks are given by

$$
\begin{aligned}
\partial_{x_k} \partial_{x_l} f &= \sum_j \frac{(c_{kj} + c_{jk})e^{y_j} e^{y_j}}{(x_k e^{y_j} + x_j e^{y_k})^2} \delta_{k,l} + \frac{(c_{kl} + c_{lk})e^{y_k} e^{y_l}}{(x_k e^{y_l} + x_l e^{y_k})^2}, \\
\partial_{y_k} \partial_{y_l} f &= -\sum_j \frac{(c_{kj} + c_{jk})x_k e^{y_j} x_j e^{y_k}}{(x_k e^{y_j} + x_j e^{y_k})^2} \delta_{k,l} + \frac{(c_{kl} + c_{lk})x_k e^{y_l} x_l e^{y_k}}{(x_k e^{y_l} + x_l e^{y_k})^2},
\end{aligned}
\tag{3.55}
$$

and off-diagonal blocks are given by

$$
\partial_{y_k} \partial_{x_l} f = \sum_j \frac{(c_{kj} + c_{jk})e^{y_k} x_j e^{y_j}}{(x_k e^{y_j} + x_j e^{y_k})^2} \delta_{k,l} - \frac{(c_{kl} + c_{lk})x_k e^{y_k} e^{y_l}}{(x_k e^{y_l} + x_l e^{y_k})^2},
\tag{3.56}
$$

$$
\partial_{x_k} \partial_{y_l} f = \partial_{y_l} \partial_{x_k} f.
$$

It is straightforward to encode the equality and inequality constraints in (3.22) into matrices A, G and vectors b, h.

$$
A = (\underbrace{0, \ldots, 0}_{n}, \underbrace{1, 0, \ldots, 0}_{n}),
\tag{3.57}
$$

$$
b = 0,
\tag{3.58}
$$

$$
G = (-I_n, 0_n),
\tag{3.59}
$$

$$
h = (0, \ldots, 0)^T
\tag{3.60}
$$

with I_n the identity and 0_n the zero matrix in $\mathbb{R}^{n \times n}$.

The Jacobian $D\Phi$ is singular because of the invariance of the objective f under a constant shift of y; this is also true for the augmented Jacobian H since the inequalities act only on x. Therefore the normal equations (3.49) cannot be formed and the search direction has to be computed from the augmented system (3.47).

The blocks of $D\Phi$ have the same sparsity pattern as the matrix $C_s = C + C^T$. This matrix is usually sparse. The augmented Jacobian differs from the original Jacobian only on the diagonal so that it is also sparse

in a situation in which C_s is sparse. The equality constraints for the reversible MLE problem do only affect the y variables, i.e. $A = (0, A_y)$. The augmented system, (3.47), can be cast into the following symmetric form,

$$
\begin{pmatrix}
H_{xx} & H_{yx} & 0 \\
H_{yx}^T & -H_{yy} & -A_y^T \\
0 & -A_y & 0
\end{pmatrix}
\begin{pmatrix}
\Delta x \\
\Delta y \\
\Delta \nu
\end{pmatrix}
=
\begin{pmatrix}
b_x \\
-b_y \\
-b_\nu
\end{pmatrix}.
\tag{3.61}
$$

The augmented system matrix, W, on the left-hand side of (3.61) is indefinite so that a symmetric indefinite factorization, [15], or the minimum residual (MINRES) method, [81], can be used to solve (3.61). If an iterative method is used, a suitable preconditioner needs to remove the ill-conditioning due to the $\Sigma = S^{-1}\Lambda$ term in H. MINRES requires a positive definite preconditioner. We use a positive definite diagonal preconditioning matrix, T, with diagonal entries,

$$
t_{ii} =
\begin{cases}
|w_{ii}| & \text{if } |w_{ii}| > 0 \\
1 & \text{else}
\end{cases}.
\tag{3.62}
$$

3.5.1 dTRAM

We can also apply the primal-dual interior-point method to the convex-concave reformulation of the dTRAM problem, (3.33). The dTRAM problem consists of a reversible MLE problem for each thermodynamic state coupled via an equality constraint. The resulting VI-mapping for dTRAM is given by the vector

$$
\Phi = (\Phi_0, \dots, \Phi_m).
$$

The entry Φ_α is the mapping for the reversible MLE problem at thermodynamic state α. Since Φ_α depends only on variables $(x^{(\alpha)}, y^{(\alpha)})$ the Jacobian of Φ_α has a block-diagonal structure

$$
D\Phi =
\begin{pmatrix}
D\Phi_0 & & \\
& \ddots & \\
& & D\Phi_m
\end{pmatrix}
$$

The matrix $D\Phi_\alpha$ is the mapping for the reversible MLE problem at thermodynamic state α. The linear inequality constraints at different α are decoupled so that G is also block diagonal,

$$
G =
\begin{pmatrix}
G_0 & & \\
& \ddots & \\
& & G_m
\end{pmatrix}.
$$

The block $G^{(\alpha)}$ is the matrix of inequality constraints at thermodynamic state α,

$$
G_\alpha = (-I_n, 0_n),
$$

and $h = 0$ is the corresponding right-hand side (RHS). The matrix for the equality constraints has the following form,

$$
A = \begin{pmatrix}
A_0 & 0 & \dots & 0 \\
A_{1,0} & A_1 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
A_{m,0} & 0 & \dots & A_m
\end{pmatrix}
$$

with $A_0 = (0, \dots, 0, 1, \dots, 0)$ the constraint matrix for the unbiased ensemble, $\alpha = 0$, and $A_\alpha = (0_n, I_n)$ the constraint matrix at condition $\alpha \neq 0$. The matrix $A_{\alpha,0} = (0_n, -I_n)$ is the coupling matrix between biased and unbiased ensemble. The corresponding RHS is

$$
b = \begin{pmatrix}
b_0 \\
\vdots \\
b_m
\end{pmatrix}
$$

with $b_0 = 0$, and $b_\alpha = (u_i^{(\alpha)})$ the vector of energy differences with respect to the unbiased condition.

The block-diagonal form of $D\Phi$ and G can be exploited for the solution of the augmented system. The block diagonal structure of $D\Phi$ and G implies a block diagonal structure for H,

$$
H = \begin{pmatrix}
H_1 & & \\
& \ddots & \\
& & H_m,
\end{pmatrix}.
\tag{3.63}
$$

The block $H_\alpha = D\Phi_\alpha + G_\alpha^T \Sigma_\alpha G_\alpha$ is the augmented Jacobian at thermodynamic state α. Using the block structure of H and A, the augmented system (3.47) can be reordered resulting in the following linear system,

$$
\begin{pmatrix}
W_0 & B_{1,0}^T & \dots & B_{m,0}^T \\
B_{1,0} & W_1 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
B_{m,0} & 0 & \dots & W_m
\end{pmatrix}
\begin{pmatrix}
\Delta\xi_0 \\
\Delta\xi_1 \\
\vdots \\
\Delta\xi_m
\end{pmatrix}
= -
\begin{pmatrix}
\tilde{b}_0 \\
\tilde{b}_1 \\
\vdots \\
\tilde{b}_m
\end{pmatrix}
\tag{3.64}
$$

The augmented system matrix at condition α is

$$
W_\alpha = \begin{pmatrix}
H_\alpha & A_\alpha^T \\
A_\alpha & 0
\end{pmatrix}.
\tag{3.65}
$$

The coupling between the biased condition and the unbiased condition is encoded in the matrix

$$
B_{\alpha,0} = \begin{pmatrix}
0 & 0 \\
A_{\alpha,0} & 0
\end{pmatrix} \quad \alpha \neq 0.
\tag{3.66}
$$

The vector $\Delta\xi_\alpha = (\Delta z_\alpha, \Delta\nu_\alpha)$ is the resulting increment for the augmented system at condition α. The vector $\tilde{b}_\alpha$ in (3.64) is given by the RHS of the augmented system at condition α,

$$\tilde{b}_\alpha = \begin{pmatrix} r_d^{(\alpha)} + G_\alpha^T \Sigma_\alpha r_{p,2}^{(\alpha)} - G_\alpha^T S_\alpha^{-1} r_c^{(\alpha)}(\mu) \\ r_{p,1}^{(\alpha)} \end{pmatrix}. \tag{3.67}$$

The arrow-shaped structure of the linear system in (3.64) allows us to apply the Schur complement method, [51, 126], to eliminate $\Delta\xi_1, \ldots, \Delta\xi_m$ and solve the following condensed system for $\Delta\xi_0$,

$$S\Delta\xi_0 = -\left(\tilde{b}_0 - \sum_{\alpha=1}^m B_{\alpha,0}^T W_\alpha^{-1} \tilde{b}_\alpha \right) \tag{3.68}$$

The Schur complement matrix is

$$S = \left(W_0 - \sum_{\alpha=1}^m B_{\alpha,0}^T W_\alpha^{-1} B_{\alpha,0} \right). \tag{3.69}$$

All other increments can be computed from $\Delta\xi_0$ via

$$\Delta\xi_\alpha = -W_\alpha^{-1} \left(\tilde{b}_\alpha + B_{\alpha,0}\Delta\xi_0 \right) \tag{3.70}$$

For a system with n states at m thermodynamic conditions the complexity for a direct factorization of the Newton system (3.43) is $\mathcal{O}(m^3 n^3)$. The Schur complement approach reduces complexity to $\mathcal{O}(mn^3)$. In addition, assembly of the Schur complement in (3.69) and solution of (3.70) can be easily parallelized.

As for the reversible MLE case, the blocks of $D\Phi_\alpha$ have the same sparsity pattern as the matrix $C_s^{(\alpha)} = C^{(\alpha)} + C^{(\alpha)T}$. The same is true for the augmented Jacobian H_α except for the diagonal. Since $C_s^{(\alpha)}$ is usually sparse we use a sparse LU method to factor the augmented system matrices W_α for $\alpha > 0$. The direct assembly of the Schur complement in (3.69) is expensive since the computation of $W_\alpha^{-1} B_{\alpha,0}$ requires $\mathcal{O}(n)$ solves.

If an iterative method is used to solve the condensed system (3.68) one would like to avoid assembly of the Schur complement S in (3.69) all together. Instead only few matrix-vector products involving S should be computed. As for the reversible MLE case, we can transform the condensed system into a symmetric indefinite form and use MINRES to obtain a solution. Obtaining a good preconditioner without explicit assembly of S is difficult. We use the probing method outlined in [16] to obtain an approximation of the diagonal of S using only few matrix-vector products. We then construct a positive definite diagonal preconditioning matrix T with entries

$$t_{ii} = \begin{cases} |\tilde{s}_{ii}| & \text{if } |\tilde{s}_{ii}| > 0 \\ 1 & \text{else} \end{cases}.$$

The entry $\tilde{s}_{ii}$ denotes the diagonal entry estimated by the probing approach.

The Schur complement based solution can also be applied to the dTRAM problem with additional constraints whenever those constraints do not couple different biasing conditions.

3.6   RESULTS

Below we report results for the primal-dual interior-point (Newton-IP) and the self consistent iteration (SC-iteration) approach to solving the reversible MLE and dTRAM problem. We compare the efficiency of both algorithms for a number of examples. Using iterative methods for the solution of the linear systems arising in the Newton-IP approach we achieve a similar scaling behavior as for the SC-iteration. We demonstrate that the Newton-IP approach offers a significant speedup for nearly all examples.

3.6.1   *Reversible MLE*

In Table 3.1 we compare the performance of the algorithm for different example data-sets. The count matrix was estimated from the full data set using the sliding-window method [84]. The tolerance indicating convergence was tol $= 10^{-12}$ for both algorithms. Both methods exhibit a subquadratic scaling in the number of states. The Newton-IP method is able to achieve a significant speed-up over the SC-iteration for all examples except for the pentapeptide data.

In Figure 3.1 we show the performance of both methods for the alanine dipeptide system with 361 states. For the SC-iteration the number of iterations required to converge to a given tolerance is very variable across different data sets. The total number of iterations required to converge deteriorates with increasing amount of input data. For the Newton-IP method the required number of iterations is consistent across all data sets. Both methods exhibit subquadratic scaling in the number of observed states.

3.6.2   *dTRAM*

In Table 3.2 we compare the performance of the Newton-IP and the SC-iteration for different examples. The count matrix was estimated from the full data set using the sliding-window method [84]. The tolerance indicating convergence was tol $= 10^{-10}$ for both algorithms. The Newton-IP method is more efficient for all three examples and achieves a dramatic speed-up (orders of magnitude). The Schur complement probing approach is successful for the alanine and the doublewell umbrella sampling example. For the multi-temperature example the Schur complement was assembled and the condensed system was solved us-

Sys.	N	N/n	Newton-IP			SC-iteration			SC/
			T	T/t	p	T	T/t	p	IP
3W	361		1.1			4.6			4.0
	2134	5.9	7.3	6.4	1.0	75.1	16.2	1.6	10.2
	8190	3.8	56.8	7.7	1.5	400.3	5.3	1.2	7.0
	29618	3.6	286.8	5.0	1.3	1076.9	2.7	0.8	3.8
Ala	292		0.7			4.2			6.3
	1059	3.6	4.2	6.4	1.4	32.3	7.8	1.6	7.6
	3835	3.6	32.2	7.6	1.6	214.0	6.6	1.5	6.6
	5826	1.5	61.8	1.9	1.6	347.7	1.6	1.2	5.6
P5	250		0.6			0.2			0.4
	500	2.0	1.2	1.9	0.9	0.6	2.4	1.3	0.5
	1000	2.0	3.6	3.0	1.6	1.0	1.8	0.9	0.3
	2000	2.0	5.4	1.5	0.6	1.3	1.3	0.4	0.2
BD	100		1.0			10.4			10.6
	200	2.0	2.1	2.1	1.1	34.1	3.3	1.7	16.3
	500	2.5	5.8	2.8	1.1	185.3	5.4	1.8	31.7
	1000	2.0	13.9	2.4	1.3	338.7	1.8	0.9	24.3

Table 3.1: Comparison of interior-point method and self-consistent iteration for the reversible MLE problem. We report the number of states N, the growth factor for states N/n (n is the number of states in the previous row), the total algorithm run time T (in seconds), the growth factor for run time T/t (t is the run time in the previous row), the scaling exponent for run time with increasing number of states p, ($T \propto N^p$), and the speedup of the Newton-IP method over the SC-iteration SC/IP. The scaling is subquadratic for both methods. The Newton-IP algorithm achieves a significant speed-up over the SC-iteration for all examples except the pentapeptide.
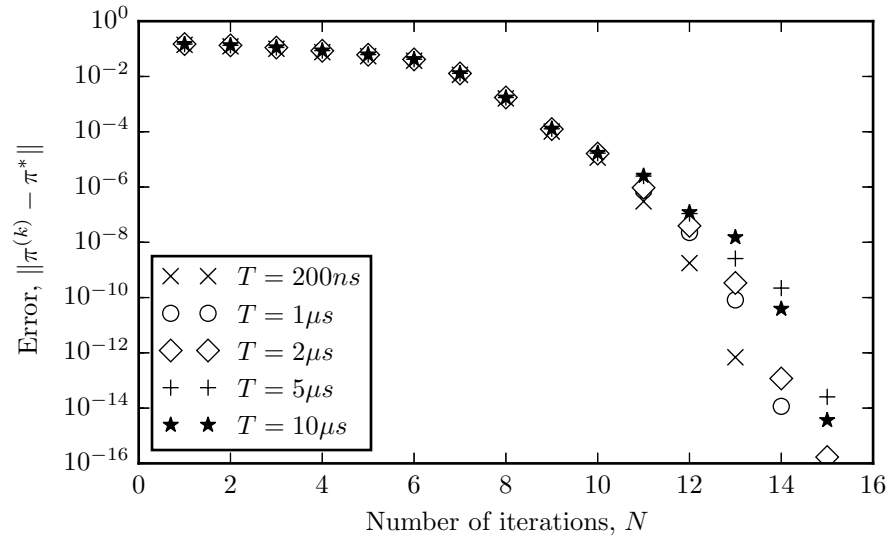
(a)
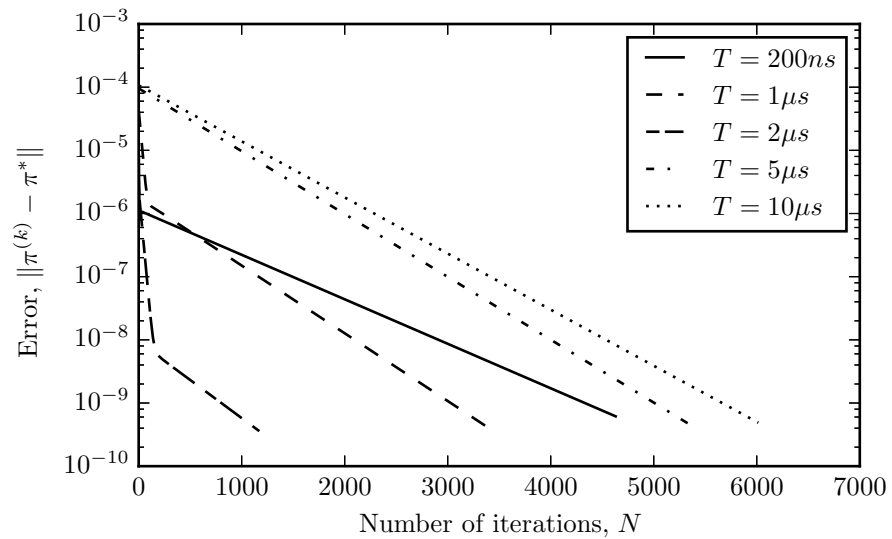
(b)

Figure 3.1: Convergence of interior-point method and self-consistent iteration for the reversible MLE problem. We show results for the Newton interior-point method, a), and the self-consistent iteration, b) for the alanine dipeptide example. Convergence is plotted for different data sets corresponding to different amounts of total simulation time. The vector π^* is a reference stationary distribution obtained from the converged Newton interior-point method. The Newton interior-point method converges superlinearly, the self-consistent iteration converges linearly. The number of required iterations is very sensitive to the input data set for the SC-iteration while the Newton-IP method is only mildly affected.
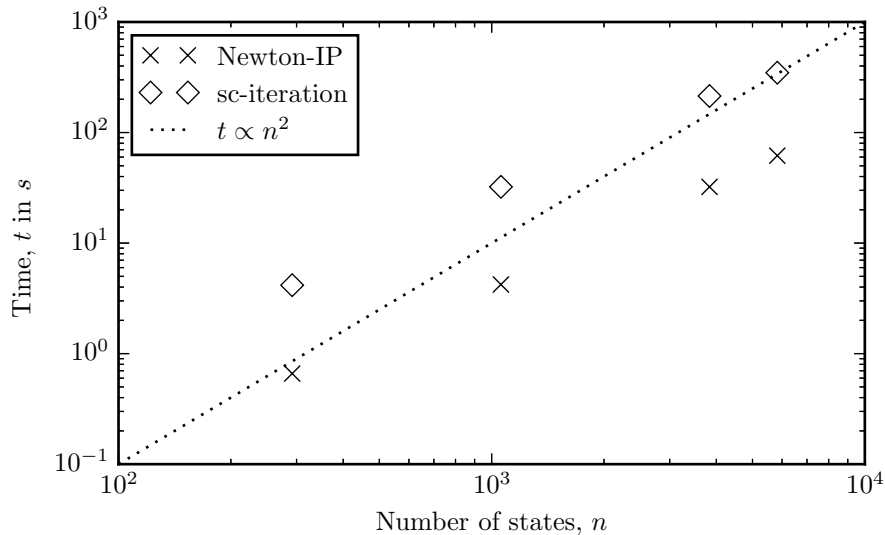
Figure 3.2: Scaling of interior-point method and self-consistent iteration for reversible MLE problem. Both methods exhibit a subquadratic scaling in the number of states. The Newton-IP method achieves a significant speed-up over the SC-iteration.

ing a direct method. For the SC-iteration method the required time to solve the multi-temperature example was very large so that computations were only carried out for two examples with a small number of states.

Both methods scale linearly in the number of thermodynamic states. The Newton-IP method with Schur complement probing scales at most quadratic in the number of states. If the Schur complement is assembled and factored by a direct method the scaling is between quadratic and cubic. The SC-iteration exhibits quadratic scaling in the number of states. The Newton-IP method achieves orders of magnitude speed-up compared to the SC-iteration for all examples.

In Figure 3.3 we show performance of the Newton-IP and SC-iteration for the doublewell umbrella-sampling example. The Newton-IP method achieves a significant speed-up (up to two orders of magnitude) over the SC-iteration.

Sys.	N	M	N/n	Newton-IP			SC-iteration			SC/IP
				T	T/t	p	T	T/t	p	
Ala	292	40		34.0			1263.9			37.2
	1521	40	5.2	202.4	6.0	1.1	66018.4	52.2	2.4	326.2
D2U	100	20		5.1			115.5			22.7
	199	20	2.0	6.4	1.3	0.3	492.9	4.3	2.1	77.1
	497	20	2.5	17.3	2.7	1.1	3258.4	6.6	2.1	188.7
	990	20	2.0	48.3	2.8	1.5	13729.7	4.2	2.1	284.4
	1978	20	2.0	193.1	4.0	2.0	59890.5	4.4	2.1	310.1
D2U	100	20		5.1			115.5			22.7
	100	40	2.0	8.3	1.6	0.7	244.5	2.1	1.1	29.3
	100	80	2.0	16.5	2.0	1.0	721.1	2.9	1.6	43.8
	100	100	1.2	20.9	1.3	1.1	1110.6	1.5	1.9	53.1
D2T	100	16		3.7			12223.2			3285.8
	200	16	2.0	10.7	2.9	1.5	50446.2	4.1	2.0	4705.8
	500	16	2.5	79.8	7.4	2.2				
	1000	16	2.0	544.5	6.8	2.8				

Table 3.2: Comparison of interior-point method and self-consistent iteration for the dTRAM problem. We report the number of states N, the number of thermodynamic state M, the growth factor for states N/n (n is the number of states in the previous row), the total algorithm run time T (in seconds), the growth factor for run time T/t (t is the run time in the previous row), the scaling exponent for run time with increasing number of states p, ($T \propto N^p$), and the speedup of the Newton IP method over the SC method SC/IP. In one case we report instead the growth factor of the number of thermodynamic states M/m (m is the number of states in the previous row) and the scaling exponent for run time with increasing number of thermodynamic states ($T \propto M^p$). Both method scale linearly in the number of thermodynamic states. The Newton-IP method with Schur complement probing (alanine, doublewell with umbrella sampling) scales at most quadratic in the number of states. If the Schur complement is assembled and factored by a direct method (doublewell with independent temperature sampling) the scaling is between quadratic and cubic. The SC-iteration exhibits quadratic scaling in the number of states. The Newton-IP method achieves orders of magnitude speed-up compared to the SC-iteration for all examples.
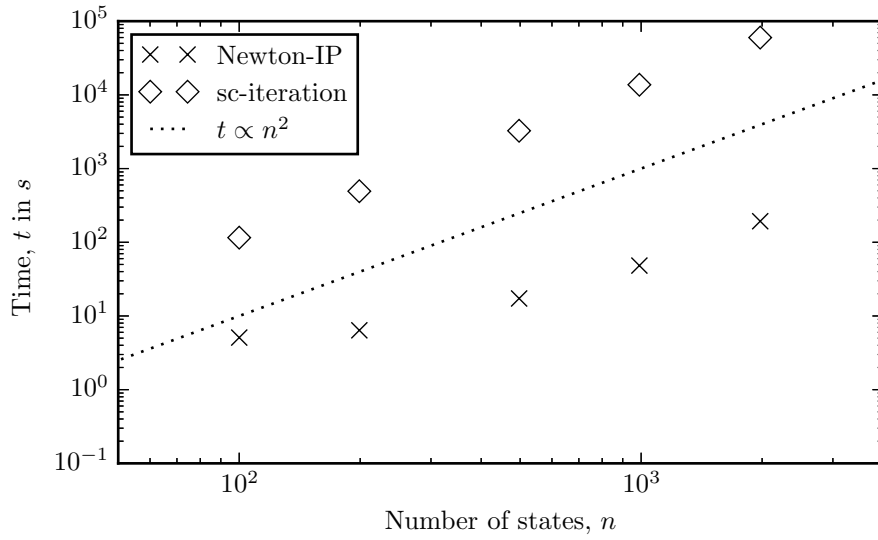
(a)
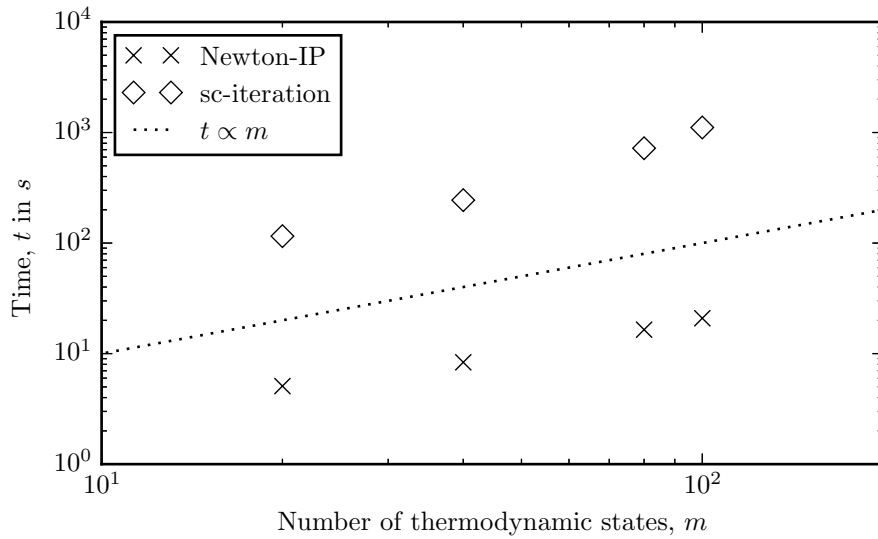
(b)

Figure 3.3: Scaling of interior-point method and self-consistent iteration for the dTRAM problem. We show results for the doublewell potential with harmonic umbrella forcing. a) Both methods exhibit quadratic scaling in the number of states, but the Newton method is up to two orders of magnitude faster then the sc iteration. b) Scaling is linear in the number of thermodynamic states for both methods.

BAYESIAN UNCERTAINTY QUANTIFICATION OF REVERSIBLE MARKOV STATE MODELS

In this chapter methods for a reliable quantitative prediction of uncertainties for reversible Markov state models are developed.

A Bayesian viewpoint is adopted and uncertainties are quantified using a posterior distribution on the set of transition matrices satisfying a detailed balance condition. A new prior ensuring that the posterior mass is concentrated around the true model even if only few observations of a rare event are present in the data is constructed.

An efficient Markov chain Monte Carlo algorithm for sampling from the posterior distribution of reversible transition matrices is developed. A variant that can incorporate *a priori* information about the stationary probabilities is also presented. Both algorithms exhibit much smaller autocorrelation times than a previous approach in [73]. They are computationally efficient and can be applied to Markov models with a large number of states. The material presented in this chapter has previously appeared in similar form in [113, 115].

4.1 THE POSTERIOR ENSEMBLE

Maximum likelihood estimation generates a point estimate, e.g. a single transition matrix that maximizes the likelihood of observing the given data. If some transition events are rare compared to the total simulation time - a typical situation for metastable systems - then one cannot expect that the estimate is a very accurate representation of the true model. In such a situation the likelihood will be 'rather flat' and observing the given data for a different model can be almost as likely as for the maximum likelihood estimate. As a result there will be an ensemble of models which is compatible with, i.e. giving similar likelihood for, the given data.

A statistical analysis of the ensemble of possible models can reveal useful information about the true model and help to quantify the uncertainty one has to attribute to estimates. For that, we need to assign probabilities to models after we have observed the data, i.e. *a posteriori*. These probabilities should also include *a priori* information and assumptions about the model. This procedure has been formalized in Bayesian statistics where the desired ensemble of models is characterized by the posterior probability distribution.

Bayes' formula relates the likelihood of an observation C given a probability model P to the posterior probability of the model given the observation,

$$\underbrace{\mathbb{P}(P|C)}_{\text{posterior}} \propto \underbrace{\mathbb{P}(P)}_{\text{prior}} \underbrace{\mathbb{P}(C|P)}_{\text{likelihood}}. \tag{4.1}$$

The posterior accounts for the uncertainty coming from a finite observation. It incorporates a priori knowledge about the quantity of interest using the prior probability.

If we are interested in an observable f that is a function of the transition matrix then we would like to compute its mean and variance from the posterior ensemble. Mean and variance of f are given by the following expectation values with respect to the posterior probability,

$$\mu[f] = \mathrm{E}[f] = \int dP\, \mathbb{P}(P|C) f(P), \tag{4.2}$$

$$\sigma^2[f] = \mathrm{E}[(f - \mu[f])^2] = \int dP\, \mathbb{P}(P|C)(f(P) - \mu[f])^2. \tag{4.3}$$

We might also be interested in the credible intervals which encompass the true value of f with some probability, such as 0.683 (1σ interval) or 0.95 (2σ interval). The posterior probability distribution of the observable, $\mathbb{P}[f]$, is of interest if we want to study the distribution of possible values of f.

As integrals (4.2) and (4.3) are high-dimensional we need to use Monte Carlo methods to approximate them. We generate a sample of transition matrices $\{P^{(k)}\}_{k=1}^N$ distributed according to the posterior and evaluate f for each sample, $P^{(k)}$. Then we approximate mean and variance by the unbiased estimators,

$$m[f] = \frac{1}{N} \sum_{k=1}^N f(P^{(k)}), \tag{4.4}$$

$$s[f] = \frac{1}{N-1} \sum_{k=1}^N (f(P^{(k)}) - m[f])^2. \tag{4.5}$$

Credible intervals are approximated by sample percentiles. The distribution of f is approximated by a histogram of the sampled values.

Before we carry out posterior inference we need to address two issues:

1. Choice of the prior: Given n Markov states, transition matrices have on the order of n^2 elements. If the choice of prior allows to populate all these elements, including those for which no transition has been observed, the probability mass of the posterior will be far away from the true model.

2. Efficiency of the sampler: For a given choice of prior and observation we need to generate samples that are distributed according to the posterior on the high-dimensional space of transition matrices in a reasonable time. This is especially problematic for reversible Markov models due to the detailed balance constraint.

4.2 SAMPLING OF NONREVERSIBLE MATRICES

Let us first illustrate the effect of prior choice on Bayesian estimation of nonreversible Markov models. A convenient functional form for the prior is the Dirichlet prior

$$\mathbb{P}(P) \propto \prod_i \prod_j p_{ij}^{b_{ij}} \tag{4.6}$$

where $B = (b_{ij})$ is a matrix of prior-counts. For this choice, the posterior is given by

$$\mathbb{P}(P|C) \propto \prod_i \left(\prod_j p_{ij}^{z_{ij}} \right). \tag{4.7}$$

The value $z_{ij} = c_{ij} + b_{ij}$ is the posterior pseudo-count. In the non-reversible case we can generate independent samples from (4.7) by drawing rows of $P^{(k)}$ independently from Dirichlet distributions $\prod_j p_{ij}^{\alpha_{ij}-1}$ with Dirichlet parameters $\alpha_{ij} = z_{ij} + 1 = c_{ij} + b_{ij} + 1$ [101], so that posterior samples can be generated efficiently.

Choosing a uniform prior, $b_{ij} = 0$, assigns equal prior probability to all entries, p_{ij}, in the posterior ensemble. But this a-priori assumption can lead to serious problems when estimating quantities for meta-stable systems.

Consider for example a birth-death chain consisting of two meta-stable sets $A = \{1, \ldots, m\}$, $B = \{m+2, \ldots, n\}$, separated by a kinetic bottleneck, the transition state $m + 1$. The transition matrix for this chain is

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & & & & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & 1-10^{-b} & 0 & 10^{-b} & & & & \\ & & & \frac{1}{2} & 0 & \frac{1}{2} & & & \\ & & & 10^{-b} & 0 & 1-10^{-b} & & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & & & & & \frac{1}{2} & \frac{1}{2} \end{pmatrix} . \tag{4.8}$$

For barrier parameter $b = 3$, $m = 50$, and $n = 101$ the expected time for hitting set B starting from state $x = 1$ is $2 \cdot 10^5$ steps.

Let us now investigate the effect that the choice of prior has on the results of the posterior inference. We illustrate the effect for a simulation of length $L = 10^7$. The true sampling distribution of the maximum likelihood estimate can be estimated with arbitrary precision by repeating the simulation many times. For the expected hitting time 10^3 repetitions led to a 90% percentile of $[1.5, 2.7] \cdot 10^5$.

In practice, we cannot afford to repeat the simulation many times but would like to approximate the true value and its statistical uncertainty from the given simulation data. Sampling the nonreversible posterior given expected counts for a chain of length $L = 10^7$ with a uniform prior, $b_{ij} = 0$, results in non-zero transition probabilities for elements p_{ij} which are zero in the true transition matrix. As a result artificial kinetic pathways circumventing the transition state appear in the posterior ensemble which leads to a dramatic underestimation of the mean first passage time. The estimate obtained from 10^3 posterior samples for the 90% credible interval is $[1.9, 2.0] \cdot 10^3$, and thus two orders of magnitude smaller than the true value $2 \cdot 10^5$.

Using the prior $b_{ij} = -1$ suggested in Ref. [77] results in a 90% credible interval, $[1.5, 2.7] \cdot 10^5$, which clearly covers the true value $2 \cdot 10^5$. This choice leads to a posterior distribution in which sampled transition matrices (p_{ij}) have the same sparsity structure as the count matrix C, i.e. $p_{ij} = 0$ if $c_{ij} = 0$. As count matrices in the present context are generally sparse, we call this prior briefly sparse prior. Apparently the sparse prior leads to consistent credible intervals covering the true value.

Fig. 4.1 shows the convergence of the 90% credible interval for the sparse and the uniform prior. The credible interval for the sparse prior envelopes the true value already given little data. To achieve consistency using the uniform prior requires simulations order of magnitudes longer than the slowest process, thus rendering inference under this prior unpractical.

Note that our prior induces a fixed sparsity structure. This concept should not be confused with other sparsity inducing priors used for example for Bayesian compressed sensing [49], where the sparsity pattern is subject to uncertainty.

4.3   SAMPLING OF REVERSIBLE MATRICES

To carry out posterior inference we need to define a prior distribution. A suitable prior distribution ensures that the posterior mass remains located around the true model to counteract the effects we have observed in the nonreversible case for a uniform prior. To define a prior distribution for the ensemble of reversible matrices we work on the space of joint (unconditional) probabilities. Then we have to generate reversible matrices distributed according to the posterior.
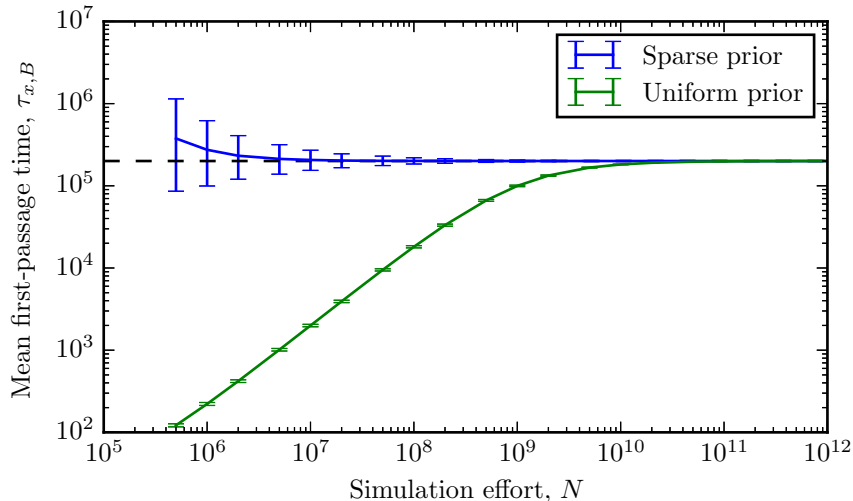
Figure 4.1: Convergence of the credible interval for nonreversible transition matrix sampling. The dashed line indicates the true value. The 90% credible interval for the sparse prior covers the true value orders of magnitude before the 90% credible interval for the uniform prior.

In the nonreversible case it was possible to generate independent samples from the posterior. Independent sampling was possible because for our choice of prior the posterior became a product of Dirichlet distributions for which drawing independent samples is well established.

There is no known method to generate independently distributed reversible transition matrices. Instead we use a Markov chain Monte Carlo (MCMC) method to generate samples from the posterior of reversible matrices. The Markov chain generates an ensemble of transition matrices, $\{P^{(k)}\}_{k=1}^{N}$, via a set of updates advancing the chain from $P^{(k)} \to P^{(k+1)}$ starting from a valid reversible matrix $P^{(0)}$. The update process ensures that each sample $P^{(k)}$ is reversible, i.e. fulfills the detailed balance condition. But, the resulting samples will be correlated.

The updates proceed by random sampling of single entries of the matrix so that our sampling procedure can be characterized as a Gibbs sampler. A collective update of all entries is difficult since the detailed balance condition together with the row normalization couples the entries in a highly nontrivial way. Finally we discuss how sampling from the conditional distributions arising in Gibbs sampling can be achieved. It turns out that independent sampling is still possible for the diagonal entries using the Beta distribution. The conditional distribution of the off-diagonal entries can be sampled via a Metropolis-Hastings chain. We show that the Gamma distribution is a good proposal distribution if the parameters are adapted to match the conditional. Finally we discuss how we can recover from heavy tails of the conditional by a simple Gaussian random walk in logspace.

4.3.1 *Prior*

We replace the Dirichlet prior (4.6) with a prior for reversible sampling. The prior ensures that the posterior mass remains located around the true model. It serves to prevent the negative effects of artificial kinetic pathways.

We perform sampling of reversible transition matrices on the space of joint probabilities

$$x_{ij} \propto \pi_i p_{ij} \tag{4.9}$$

For a reversible matrix we have $x_{ij} = x_{ji}$. We restrict use to the set of independent variables with $i \geq j$ and keep them normalized to one,

$$\sum_{i \geq j} x_{ij} = 1. \tag{4.10}$$

We define a prior for reversible sampling on the set of X matrices rather than on P,

$$\mathbb{P}(X) \propto \prod_{i \geq j} x_{ij}^{b_{ij}-1}. \tag{4.11}$$

The posterior for reversible sampling is then given as

$$\mathbb{P}(X|C) \propto \prod_{i \geq j} x_{ij}^{b_{ij}} \prod_{i,j} \left(\frac{x_{ij}}{\sum_k x_{ik}} \right)^{c_{ij}}. \tag{4.12}$$

Below we will first discuss how to sample from (4.12) using general prior counts, b_{ij}. Then, we will consider the specific choice $b_{ij} = -1$ and show that this choice has similar properties as the sparse prior in the nonreversible case.

4.3.2 *Algorithm*

We use the following shorthand notation $\sum_k x_{ik} = x_i$, $\sum_{k \neq j} x_{ik} = x_{i,-j}$, and $\sum_{i \geq j} x_{ij} = x_0$ for sums over matrix entries.

Unfortunately the normalization condition enforces us to update all coordinates together. Furthermore any single coordinate is restricted to lie in $[0, 1)$. To circumvent this problem and enable sampling of single coordinates we transform the posterior into new coordinates via the following change of variables for fixed k, l,

$$v_{ij} = \frac{x_{ij}}{1 - x_{kl}}. \tag{4.13}$$

As a result v_{kl} can take arbitrary positive values, $v_{kl} \geq 0$, and the resulting normalization condition is independent of the current coordinate,

$$\sum_{i \geq j/(k,l)} v_{ij} = 1. \tag{4.14}$$

The elements of P are then simply given by

$$p_{ij} = \frac{v_{ij}}{\sum_k v_{ik}}. \tag{4.15}$$

As we show in Section B.1.1, the conditional distribution for the diagonal element v_{kk} is

$$\gamma(v'_{kk}|V) \propto (1 + v'_{kk})^{-(b_0 + \frac{n(n+1)}{2})} v_{kk}'^{c_{kk} + b_{kk} - 1} (v_{k,-k} + v'_{kk})^{-c_k}. \tag{4.16}$$

The conditional distribution for the off-diagonal element v_{kl} is

$$\gamma(v'_{kl}|V) \propto (1 + v'_{kl})^{-(b_0 + \frac{n(n+1)}{2})} v_{kl}'^{c_{kl} + c_{lk} + b_{kl} - 1} \\ (v_{k,-l} + v'_{kl})^{-c_k} (v_{l,-k} + v'_{kl})^{-c_l}. \tag{4.17}$$

We can perform Gibbs sampling as follows. Pick an entry (k, l) with $k \geq l$. Normalize all elements except v_{kl} so that their sum is unity. Generate v'_{kl} according to (4.16) for $k = l$ and according to (4.17) otherwise. Proceed to next element.

In practice one cannot draw independent samples from the conditional in (4.17). Therefore one uses a proposal density $q(v'_{kl}|V)$ to propose an update. The proposed update is accepted with probability $\min\{1, p_{\text{acc}}\}$ with

$$p_{\text{acc}} = \frac{\gamma(v'_{kl}|V)q(v_{kl}|V')}{\gamma(v_{kl}|V')q(v'_{kl}|V)}. \tag{4.18}$$

This approach will work for any choice of prior counts. Its efficiency will crucially depend on the choice of proposal density.

We use the sparse prior, $b_{ij} = -1$, with the hope to achieve similar good results as in the nonreversible case. For this choice the $(1 + v'_{kl})^{-(b_0 + \frac{n(n+1)}{2})}$ prefactor drops out and the conditionals are scale invariant, $\gamma(v'_{kl}|V) \propto \gamma(\alpha v'_{kl}|\alpha V')$ for any $\alpha > 0$. Since the transformation from V to P is also scale invariant we can drop the normalization condition completely and set V' as new sample $V^{(k)}$ once we accept it.

The conditional density degenerates to a point probability at zero if $c_{kl} + c_{lk} = 0$, which implies that $b_{ij} = -1$ encodes the *a priori* belief that any transition for which neither the forward direction nor the backward direction has been observed in the data has zero probability in the posterior ensemble. Thus, this prior enforces P to have the same sparsity structure as the count matrix, like the choice $b_{ij} = -1$ for nonreversible sampling.

As shown in Section B.1.2 sampling of the diagonal elements can be achieved by sampling a transformed variable from the Beta distribution

$$s \sim \text{Beta}(c_{kk}, c_{k,-k})$$
$$v'_{kk} = \frac{s}{1 - s} v_{k,-k} \tag{4.19}$$

For $k \neq l$ we have no way to draw independent samples, but $\gamma(v'_{kl}|V)$ can be well approximated by a Gamma distribution matching the maximum point and the second derivative at the maximum. A Gamma distribution can be efficiently sampled and we use it as a proposal distribution for Metropolis Hastings chain on the coordinate v_{kl},

$$q(v'_{kl}|V) = \mathrm{Gamma}(v'_{kl}|\alpha, \beta). \tag{4.20}$$

In Section B.1.3 we show that parameters α, β achieving a match are given by $\alpha = -h\bar{v}$, $\beta = -h\bar{v}^2$ with

$$\bar{v} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{4.21a}$$

$$h = -\frac{c_{kl} + c_{lk}}{\bar{v}^2} + \frac{c_k}{(v_{k,-l} + \bar{v})^2} + \frac{c_l}{(v_{l,-k} + \bar{v})^2} \tag{4.21b}$$

$$a = c_k + c_l - c_{kl} - c_{lk} \tag{4.21c}$$

$$b = c_k v_{l,-k} + c_l v_{k,-l} - (c_{kl} + c_{lk})(v_{k,-l} + v_{l,-k}) \tag{4.21d}$$

$$c = -(c_{kl} + c_{lk})v_{k,-l}v_{l,-k} \tag{4.21e}$$

and that the acceptance probability for the step is $\min\{1, p_{acc}\}$ with

$$\log p_{\mathrm{acc}} = \beta(v'_{kl} - v_{kl}) + (c_{kl} + c_{lk} - \alpha) \log \frac{v'_{kl}}{v_{kl}}$$
$$-c_k \log \frac{v_{k,-l} + v'_{kl}}{v_{k,-l} + v_{kl}} - c_l \log \frac{v_{l,-k} + v'_{kl}}{v_{l,-k} + v_{kl}}. \tag{4.22}$$

If the current value of v_{kl} is in one of the heavy tails of the distribution $\gamma(v'_{kl}|V)$, the acceptance probability for the Gamma proposal can be very small and the Markov chain can get stuck. To avoid this problem we perform a second sampling step. We perform a Gaussian random walk on $\log v'_{kl}$,

$$q(\log v'_{kl}|V) = \mathcal{N}(\log v_{kl}, 1),$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. We accept the random walk proposal with $\min\{1, p_{\mathrm{acc}}\}$. We show in Section B.1.4 that

$$\log p_{\mathrm{acc}} = (c_{kl} + c_{lk}) \log \frac{v'_{kl}}{v_{kl}} - c_k \log \frac{v_{k,-l} + v'_{kl}}{v_{k,-l} + v_{kl}} - c_l \log \frac{v_{l,-k} + v'_{kl}}{v_{l,-k} + v_{kl}}, \tag{4.23}$$

In summary the proposed Algorithm 4 is a Metropolis within Gibbs MCMC algorithm with adapted proposal probabilities for each Gibbs sampling step.

Algorithm 4 : Reversible sampling algorithm

Input : C, $V^{(j)}$
Output : $V^{(j+1)}$
for *all indexes* (k, l) *with* $k \geq l$ *and* $c_{kl} + c_{lk} > 0$ **do**
    if $k = l$ **then**
        Sample $v_{kk}^{(j+1)}$ according to (4.19)
    else
        Compute parameters α, β, cf. (4.21)
        $v_{kl}' \sim \text{Gamma}(\alpha, \beta)$
        Compute p_{acc} from (4.22)
        $v_{kl}^{(j+1)} = v_{kl}'$ with probability $\min\{1, p_{\text{acc}}\}$
        $v_{lk}^{(j+1)} = v_{kl}^{(j+1)}$
        $\log v_{kl}' \sim \mathcal{N}(\log v_{kl}^{(j+1)}, 1)$
        Compute p_{acc} from (4.23)
        $v_{kl}^{(j+1)} = v_{kl}'$ with probability $\min\{1, p_{\text{acc}}\}$
        $v_{lk}^{(j+1)} = v_{kl}^{(j+1)}$
    end
end

4.3.3   *Validation*

In Figure 4.2 we compare the histogram generated by Algorithm 4 with analytical values for the posterior density for the following count matrix

$$C = \begin{pmatrix} 5 & 2 \\ 3 & 10 \end{pmatrix} \tag{4.24}$$

We can compare the histogram to the nonreversible posterior since any 2×2 transition matrix is reversible. For the nonreversible posterior we use Dirichlet prior counts $b_{ij} = -1$. We expect the analytical and sampled densities to be equal. The histogram generated by Algorithm 4 is indeed distributed according to the analytical posterior.

4.3.4   *Application*

Here, we apply Algorithm 4 to molecular dynamics simulation data of alanine dipeptide. We show that the sparse prior, (4.12) with prior counts $b_{ij} = -1$, leads to credible intervals containing the reference value for different observables.

The system was simulated on GPU-hardware using the OpenMM simulation package [27] using the *amber99sb-ildn* forcefield [62] and the *tip3p* water model [50]. The cubic box of length $2.7nm$ contained a total of 652 solvent molecules. We used Langevin equations at $T = 300K$ with a time-step of $2fs$.
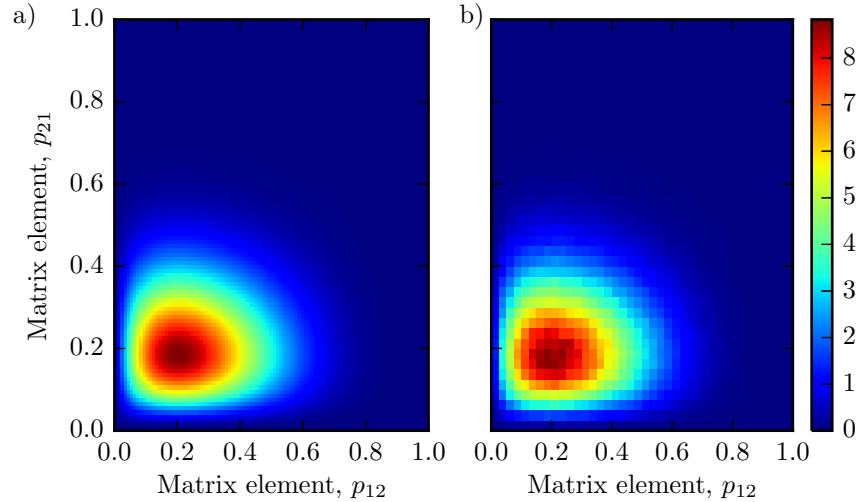
Figure 4.2: Posterior density and histogram for sampling of reversible matrices. We show results for a 2×2 count matrix. The sample histogram b) obtained from Algorithm 4 approximates the analytical density a).

We apply the sampling algorithm to $1\mu s$ of simulation data and use the MLE for $10\mu s$ of simulation data as reference value. The discretization is a regular 20×20 grid for the ϕ and ψ dihedral angles. Transition counts are obtained by sampling one count per lag time. We estimate mean, 90% credible interval, and histogram from a sample of $N = 10^5$ reversible matrices for all observables. As observables we use largest relaxation timescales t_i and expected hitting times $\tau(A \rightarrow B)$. We compute t_2, t_3, t_4, and $\tau(A \rightarrow B)$ for three transitions between metastable sets, $C_5 \rightarrow C_7^{ax}$, $C_5 \rightarrow \alpha_L$ and $C_5 \rightarrow \alpha_R$.

In Figure 4.3 we show histograms for time scales and expected hitting times computed from the posterior sample. Reference values and sample mean are indicated by dashed and dotted lines. The 90% credible intervals, in gray, contain the reference value for all observables. In Table 4.1 we report the reference value together with sample mean, sample standard deviation, error of the mean and autocorrelation time. The reference value can be found within two standard deviations from the mean for all observables.

4.3.5   *Efficiency*

We compare the efficiency of Algorithm 4 with a previous algorithm [73]. The previous algorithm also uses a Gibbs sampling strategy, but instead of posterior adapted densities it uses uniform densities to propose updates. As a measure of efficiency we use autocorrelation times for the largest relaxation timescale t_2. We also report acceptance probabilities for the current and the previous algorithm. We use the alanine dipeptide data described in the previous section [CITE]. Autocorrela-
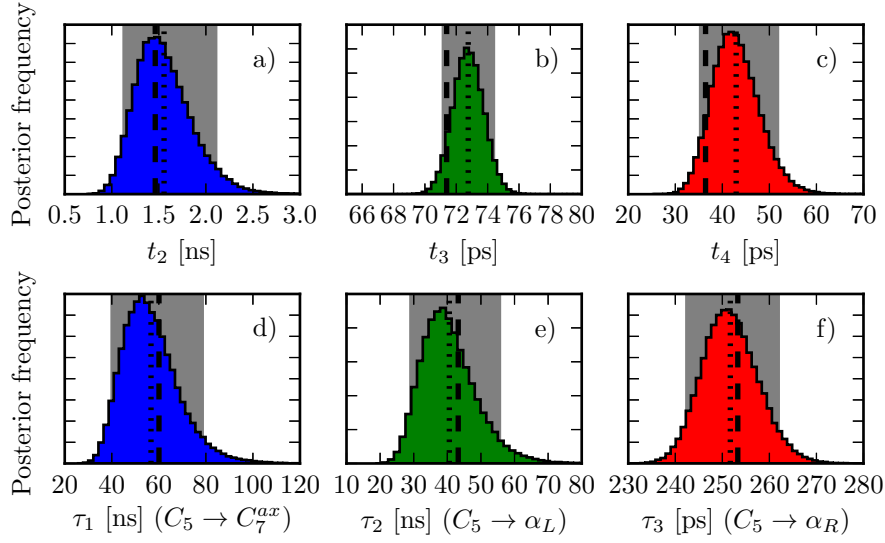
Figure 4.3: Histograms for a sample of reversible matrices. Observables are computed for a posterior sample generated by Algorithm 4. Figures a)-c) show implied time-scales, and figures d)-f) show expected hitting times. Dashed lines indicate the reference value, dotted lines indicate the sample mean. The 90% credible intervals are the shaded regions in gray. The reference value is compatible with the posterior sample (credible interval) in all cases.

	$\hat{x}$	m	s	ϵ	t_{corr}
t_2	1462	1556	303	19.00	197
t_3	71	73	1	0.01	10
t_4	36	43	5	0.06	7
$\tau(C_5 \to C_7^{ax})$	60.4	56.7	12.0	0.77	202
$\tau(C_5 \to \alpha_L)$	43.6	40.6	8.3	0.53	206
$\tau(C_5 \to \alpha_R)$	0.253	0.250	0.005	0.0004	218

Table 4.1: Posterior estimates for sampling of reversible matrices. The table contains values for relaxation time scales t_i (in ps) and and expected hitting times (in ns). Mean m, standard deviation s, error of the mean ϵ and autocorrelation time t_{corr} were estimated using $N = 10^5$ samples from the reversible posterior. The reference value $\hat{x}$ is contained within two standard deviations of the mean for all observables.

algorithm	n	p_{offdiag}	p_{diag}	t_{corr}
previous	233	0.216	0.011	1088.1
	1108	0.271	0.005	3241.9
current	233	0.994	1.0	194.7
	1108	0.995	1.0	242.6

Table 4.2: Acceptance probabilities and autocorrelation times for sampling of reversible matrices. We show results for the current method, Algorithm 4, and and a previous method, [73]. Acceptance probability for sampling of diagonal and off-diagonal elements, p_{offdiag}, p_{diag}, and autocorrelation time for the slowest relaxation timescale t_{corr} were estimated from $N = 10^6$ posterior samples for systems with n states. Acceptance probabilities for the current method are nearly optimal. p_{diag} is extremely small for the previous method. The autocorrelation times for the current algorithm are much smaller than for the previous algorithm.

tion functions and autocorrelation times are computed from $N = 10^6$ samples. Two differently fine discretizations were used, resulting in a small and a large state space with $n = 233$ and $n = 1108$ states.

Figure 4.4 shows autocorrelation functions of t_2 for both algorithms. The autocorrelation functions for Algorithm 4 decay much faster than those for the algorithm in [73]. Correlation between samples increases significantly with increasing dimension for the previous algorithm while the current algorithm is only mildly affected.

In Table 4.2 we report acceptance probabilities for the update of diagonal and off-diagonal matrix elements for both algorithms. The posterior adapted proposal steps in the current algorithm results in very high acceptance probabilities for off-diagonal updates. Diagonal elements can be sampled independently so that no rejection is necessary. The previous algorithm has a very low acceptance probability for the diagonal updates and the acceptance probability for the off-diagonal element is significantly lower than for the current algorithm. Autocorrelation times for the current algorithm are more than a factor 5 shorter for the small (233 state) system and more than a factor 13 shorter for the large (1108 state) system. The autocorrelation time increases significantly with increasing dimension for the previous algorithm while the current algorithm is only mildly affected.

The autocorrelation time increases only mildly for matrices of increased dimension for the current algorithm, there is a large increase for the previous algorithm.

The diagonal update is the only step that modifies the stationary vector in the previous algorithm. As a result it is mixing very poorly. The smaller autocorrelation time for the current algorithm indicates a higher efficiency compared to the previous algorithm. The mild dependence on state space dimension indicates that it will also be useful for large Markov models.
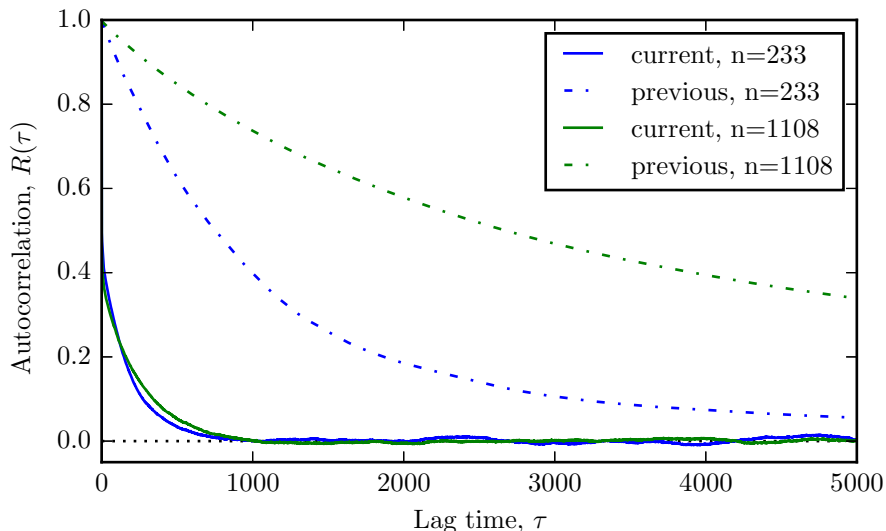
Figure 4.4: Autocorrelation functions for sampling of reversible matrices. We show results for the current method, Algorithm 4, (solid lines) and a previous method, [73], (dashed lines) for systems with $n = 233$ states (blue) and $n = 1108$ states (green). For the previous algorithm the autocorrelation time increases significantly for the larger system while the autocorrelation time for the current algorithm is only mildly affected.

4.4 SAMPLING WITH A FIXED STATIONARY VECTOR

Our approach to sampling of reversible matrices with a given stationary vector is very similar to the one outlined above for the sampling of reversible matrices. We use the same prior, (4.11), as in the reversible case. In contrast to the reversible case, we need to regularize some of the diagonal prior counts and we need to ensure that the initial matrix has positive diagonal entries to implement a working Gibbs sampler.

4.4.1 *Prior*

As before we use variables $x_{ij} = \pi_i p_{ij}$. In contrast to the previous algorithm, π is not changed by the Monte Carlo update but remains fixed. We replace the normalization condition (4.10) by a constraint for each row,

$$\sum_j x_{ij} = \pi_i. \tag{4.25}$$

All x_{kl} in the strictly lower triangle $(k > l)$ are used as independent variables for Gibbs sampling. Given a valid X matrix, i.e. positive, symmetric and obeying (4.25), we update it via

$$x_{kl} \to x'_{kl} \tag{4.26a}$$
$$x_{kk} \to x_{kk} + (x'_{kl} - x_{kl}) \tag{4.26b}$$
$$x_{lk} \to x'_{kl} \tag{4.26c}$$
$$x_{ll} \to x_{ll} + (x'_{lk} - x_{lk}). \tag{4.26d}$$

(4.26b) and (4.26d) ensure that the normalization condition (4.25) holds for the new sample, while (4.26c) restores the symmetry.

We sample x'_{kl} from the conditional distribution,

$$\gamma(x'_{kl}|X) \propto (x'_{kl})^{c_{kl}+c_{lk}+b_{kl}} \left(\Delta_{kl} - x'_{kl}\right)^{c_{kk}+b_{kk}} \left(\Delta_{lk} - x'_{kl}\right)^{c_{ll}+b_{ll}}, \tag{4.27}$$

with parameter $\Delta_{kl} = x_{kk} + x_{kl}$.

We have seen that a correct choice of prior parameters was essential in order to successfully apply the posterior sampling for meta stable systems. As in the reversible case we will use $b_{kl} = -1$ for $k > l$ to enforce $x_{kl} = 0$ whenever $c_{kl} + c_{lk} = 0$. For $c_{kk} > 0$ we also use the prior $b_{kk} = -1$.

In the case $c_{kk} = 0$ we can not set $b_{kk} = -1$. For $c_{kk} = 0$ and $b_{kk} = -1$ the conditional distribution (4.27) will degenerate to a point mass at Δ_{kl} so that $x'_{kk} = 0$ with probability one. As a result the k-th row and column of X will remain fixed in the sampling process and the resulting chain will not be ergodic. If we choose $b_{kk} = 0$ for all k with $c_{kk} = 0$ then $x_{kk} > 0$ for all such k and the posterior expectation will be positive.

If the maximum likelihood estimate, $\hat{x}_{kk}$, is zero for a k with $c_{kk} = 0$ then the posterior expectation should also be small to minimize self-transitions that are artificially stabilized by the sampling process. In that case we propose to regularize the prior using a small positive parameter, $\epsilon > 0$, and set $b_{kk} = -1 + \epsilon$. Finally, if $\hat{x}_{kk} > 0$ for a k with $c_{kk} = 0$ we set $b_{kk} = 0$ so that the expectation of (4.27) matches the MLE of the one-dimensional likelihood function for x_{kl}, cf. Section B.2.1.

In summary, we select the prior for reversible sampling with fixed stationary vector as

$$b_{kk} = \begin{cases} -1 & \text{if } c_{kk} > 0 \\ -1 + \epsilon & \text{if } \hat{p}_{kk} = 0, \ c_{kk} = 0 \\ 0 & \text{if } \hat{p}_{kk} > 0, \ c_{kk} = 0 \end{cases} \tag{4.28}$$

In addition we need to ensure that the Markov chain is started from an initial state $X^{(0)}$ with strictly positive entries, $x_{kk}^{(0)} > 0$, so that the update (4.26) can be performed for each row.

4.4.2 *Algorithm*

We now investigate how to efficiently sample from the conditional (4.27). We assume, without loss of generality, that $x_{kk} \leq x_{ll}$ and transform $x'_{kl} \in (0, \Delta_{kl})$ into a new variable $v' \in [0, +\infty)$ via

$$v' = \frac{x'_{kl}}{\Delta_{kl} - x'_{kl}}. \tag{4.29}$$

In Section B.2.2 we show that the transformed conditional is

$$\gamma(v'|X) \propto (v')^{a_1} (r+v)^{a_3} (1+v)^{-(a_1+a_2+a_3+2)} \tag{4.30}$$

with parameters

$$r = \frac{s}{s-1} \tag{4.31a}$$

$$s = \frac{\Delta_{lk}}{\Delta_{kl}} \tag{4.31b}$$

$$a_1 = c_{kl} + c_{lk} + b_{kl} \tag{4.31c}$$

$$a_2 = c_{kk} + b_{kk} \tag{4.31d}$$

$$a_3 = c_{ll} + b_{ll} \tag{4.31e}$$

We can again use a Gamma proposal density to approximate the conditional $\gamma(v'|X)$. In Section B.2.3 we show that the parameters for the Gamma distribution are $\alpha = -h\bar{v}$, $\beta = -h\bar{v}^2$ with

$$\bar{v} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{4.32a}$$

$$h = -\frac{a_1}{\bar{v}^2} - \frac{a_3}{(r+\bar{v})^2} + \frac{a_2}{(1+\bar{v})^2} \tag{4.32b}$$

$$a = a_2 + 1 \tag{4.32c}$$

$$b = a_2 - a_1 + \frac{a_2 + a_3 + 1}{s - 1} \tag{4.32d}$$

$$c = r(a_1 + 1) \tag{4.32e}$$

and that the proposed value v' has to be accepted with probability $\min\{1, p_{\mathrm{acc}}\}$ where

$$\log p_{\mathrm{acc}} = \beta(v' - v) + (a_1 + 1 - \alpha) \log \frac{v'}{v} + a_3 \log \frac{r + v'}{r + v} \\ - (a_1 + a_2 + a_2 + 2) \log \frac{1 + v'}{1 + v}. \tag{4.33}$$

In addition, to prevent the update algorithm from getting stuck in the tails of $\gamma(v'|X)$ we also perform a Gaussian random walk on $\log v'$. In Section B.2.4 we show that the proposed step needs to be accepted with probability $\min\{1, p_{\mathrm{acc}}\}$ where

$$\log p_{\mathrm{acc}} = (a_1 + 1) \log \frac{v'}{v} + a_3 \log \frac{r + v'}{r + v} \\ - (a_1 + a_2 + a_3 + 2) \log \frac{1 + v'}{1 + v}. \tag{4.34}$$

The proposed Algorithm 5 for sampling of reversible transition matrices with fixed stationary vector can again be characterized as a Metropolis within Gibbs MCMC algorithm with adapted proposal probabilities.

Algorithm 5 : Reversible sampling algorithm with fixed stationary vector

---

Input : C, π, $X^{(j)}$
Output : $X^{(j+1)}$
$X^{(j+1)} = X^{(j)}$
for *all indexes* (k, l) *with* $k > l$ *and* $c_{kl} + c_{lk} > 0$ **do**
    if $x_{kk} > x_{ll}$ **then**
        Swap k with l
    end
    Compute parameters α, β using (4.32)
    $v' \sim \mathrm{Gamma}(\alpha, \beta)$
    $x'_{kl} = \Delta_{kl} \frac{v'}{1+v'}$
    Compute p_{acc} according to (4.33)
    Update $X^{(j+1)}$ according to (4.26) with probability
    $\min\{1, p_{\mathrm{acc}}\}$
    $X^{(j)} = X^{(j+1)}$
    $\log v' \sim \mathcal{N}(\log v, 1)$
    Compute p_{acc} according to (4.34)
    $x'_{kl} = \Delta_{kl} \frac{v'}{1+v'}$
    Update $X^{(j+1)}$ according to (4.26) with probability
    $\min\{1, p_{\mathrm{acc}}\}$
end

4.4.3 *Validation*

In Figure 4.5 we compare the histogram generated by Algorithm 5 with analytical values for the posterior density. We use the 2×2 count matrix in (4.24) and the stationary vector $\pi = (0.25, 0.75)^\top$. The detailed balance condition with fixed stationary vector enforces a linear dependency between the transition matrix element p_{12} and p_{21}. The resulting posterior is therefore restricted to the line $\pi_1 p_{12} = \pi_2 p_{21}$. The projection onto p_{12} contains the full information about the one-dimensional posterior. The histogram generated by Algorithm 5 is distributed according to the analytical posterior.

4.4.4 *Application*

We apply Algorithm 5 to molecular dynamics simulation data of alanine dipeptide. We show that the sparse prior, (4.12) with prior counts in
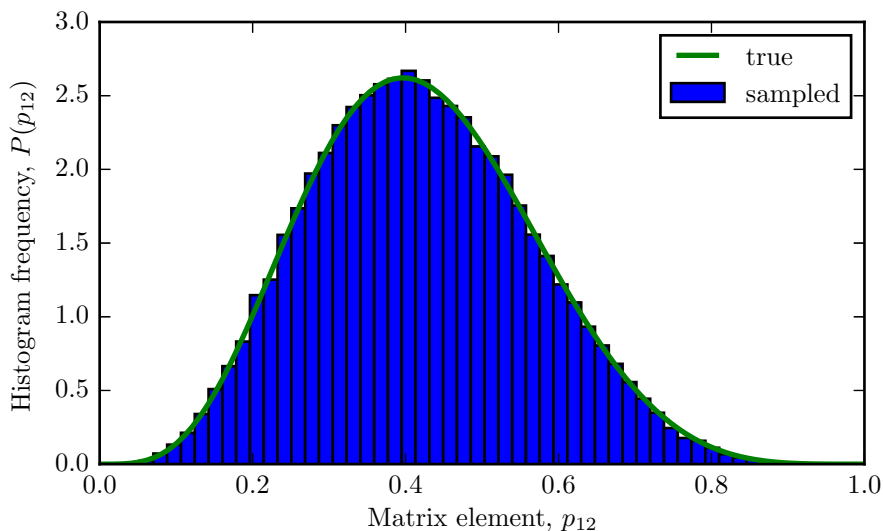
Figure 4.5: Posterior density and histogram for sampling of reversible matrices with fixed stationary vector. We show results for a 2×2 count matrix. The sample histogram obtained from Algorithm 5 approximates the analytical density.

(4.28), leads to credible intervals containing the reference value for different observables.

We use the setup described in Section 4.3.4. For the computation of the reference value we use the MLE transition matrix of the full simulation data reversible with respect to the input stationary distribution for the posterior sampling.

The stationary distribution π_i was approximated by the relative frequency of states which can be computed from the count matrix,

$$\pi_i = \frac{\sum_k c_{ik}}{\sum_{j,k} c_{jk}}. \tag{4.35}$$

It should be noted that a more accurate approximation for the stationary vector can be obtained from enhanced sampling simulations targeted at rapidly generating a good estimate of the equilibrium probabilities alone. See [113] for methods and applications that combine MD simulations and enhanced sampling simulations in order to efficiently compute rare-event kinetics.

In Figure 4.6 we show histograms for time scales and expected hitting times computed from the posterior sample. Reference values and sample mean are indicated by dashed and dotted lines. The 90% credible intervals, in gray, contain the reference value for all observables. In Table 4.3 we report the reference value together with sample mean, sample standard deviation, error of the mean, and autocorrelation time. The reference value can be found within two standard deviations from the mean for all observables.

The additional constraint imposed by fixing the stationary distribution is clearly reflected in smaller standard deviations for all shown ob-
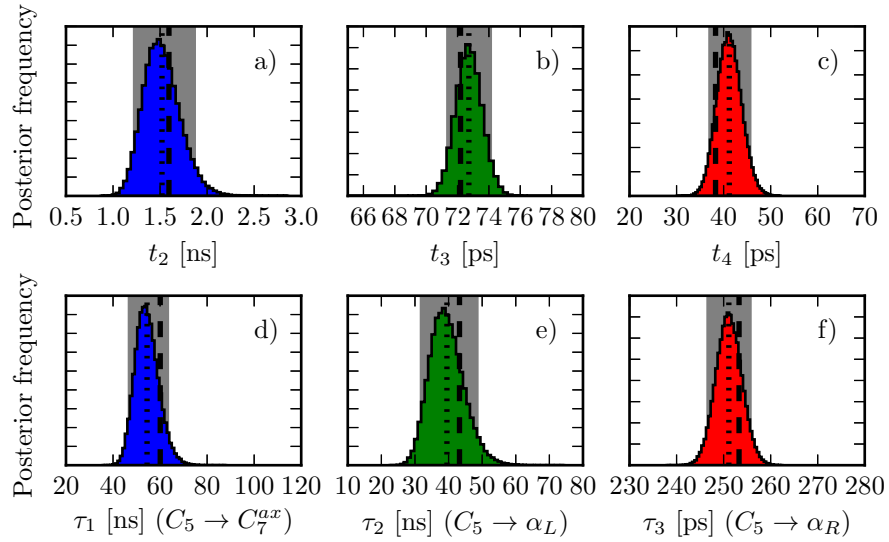
Figure 4.6: Histograms for a sample of reversible matrices with fixed stationary vector. Observables are computed for a posterior sample generated by Algorithm 5. Figures a)-c) show implied time-scales and figures d)-f) show expected hitting times. Dashed lines indicate the reference value, dotted lines indicate the sample mean. The 90% credible intervals are the shaded regions in gray. The reference value is compatible with the posterior sample (credible interval) in all cases.

servables compared to the reversible case, cf. Table 4.1. This indicates that the stationary distribution can be a valuable source of information for inference about kinetic quantities.

4.4.5 *Efficiency*

We compare the efficiency of Algorithm 5 with a previous algorithm [73]. As a measure of efficiency we use autocorrelation times for the largest relaxation timescale t_2. We also report acceptance probabilities for the current and the previous algorithm. The setup is identical to the one outlined in Section 4.3.5, the stationary distribution is computed using (4.35).

In Figure 4.7 we show autocorrelation functions of t_2 for both algorithms. The autocorrelation functions for Algorithm 5 decay much faster than those for the algorithm in [73]. Correlation between samples increases significantly with increasing dimension for the previous algorithm while the current algorithm is only mildly affected.

Table 4.4 compares acceptance probabilities and autocorrelation times for the two algorithms. For sampling with fixed stationary vector there is no sampling step for the diagonal elements. Although the average acceptance probabilities are only a factor of 3-4 better for our new algorithm, the autocorrelation times are decreased by a factor 35 for the small system (233 states) and over a factor 300 for the large system

	$\hat{x}$	m	s	ϵ	t_{corr}
t_2	1594	1520	196	0.6	1
t_3	72	73	1	0.003	1
t_4	38	41	3	0.01	1
$\tau(C_5 \rightarrow C_7^{ax})$	56.0	54.5	5.0	0.02	1
$\tau(C_5 \rightarrow \alpha_L)$	41.5	39.5	5.1	0.02	1
$\tau(C_5 \rightarrow \alpha_R)$	0.249	0.251	0.003	0.00001	1

Table 4.3: Posterior estimates for sampling of reversible matrices with fixed stationary vector. The table contains values for relaxation time scales t_i (in ps) and and expected hitting times (in ns). Mean m, standard deviation s, error of the mean ϵ and autocorrelation time t_{corr} were estimated using $N = 10^5$ samples from the reversible posterior. The reference value $\hat{x}$ is contained within two standard deviations of the mean for all observables.

algorithm	n	p_{offdiag}	t_{corr}
previous	233	0.175	72.096
	1108	0.230	> 1000
current	233	0.752	2.893
	1108	0.706	3.157

Table 4.4: Acceptance probabilities and autocorrelation times for sampling of reversible matrices with fixed stationary vector. We show results for the current method, Algorithm 5, and and a previous method, [73]. Acceptance probability for sampling of off-diagonal elements p_{offdiag} and autocorrelation time for the slowest relaxation timescale t_{corr} were estimated from $N = 10^6$ posterior samples for systems with n states. For the previous algorithm the autocorrelation time increases significantly for the larger system while the autocorrelation time for the current algorithm is only mildly affected. Autocorrelation function for the large system for the previous method did not converge.

(1108 states). Again, the autocorrelation time increases significantly with increasing dimension for the previous algorithm while the current algorithm is only mildly affected.

The smaller autocorrelation time for the current algorithm indicates a higher efficiency compared to the previous algorithm. The mild dependence on state space dimension indicates that it will also be useful for large Markov models.

4.5 INFERENCE USING AN UNCERTAIN STATIONARY VECTOR

We discuss how *a priori* information about the stationary vector can be used in situations in which the stationary vector is not known with absolute certainty. The situation we have in mind is one in which the stationary vector can be estimated from another independent set of enhanced sampling simulations that cannot be used to estimate the
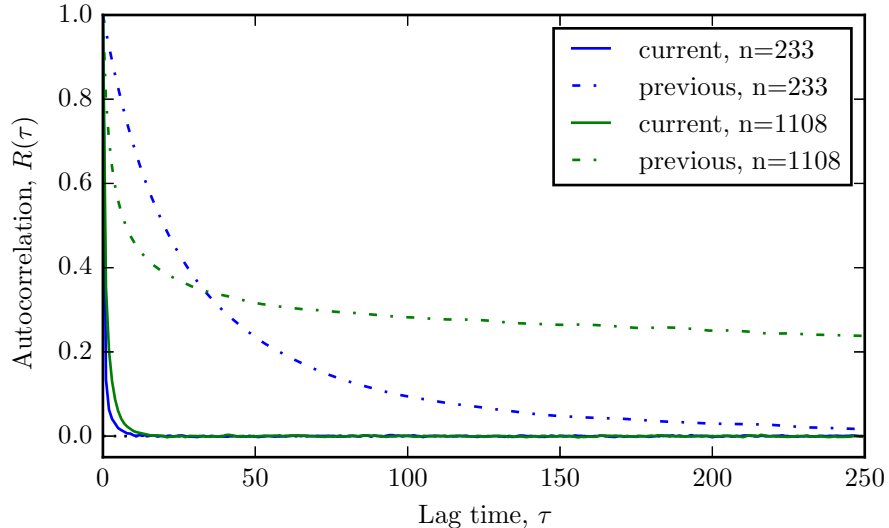
Figure 4.7: Autocorrelation function for reversible sampling with fixed stationary vector. We show results for the current method, Algorithm 5, (solid lines) and a previous method, [73], (dashed lines) for systems with $n = 233$ states (blue) and $n = 1108$ states (green). For the previous algorithm the autocorrelation time increases significantly for the larger system while the autocorrelation time for the current algorithm is only mildly affected.

transition matrix. The stationary vector estimate is uncertain since we use a finite amount of data. This finite sampling error should be accounted for during posterior inference (of reversible matrices).

From a Bayesian viewpoint we have to combine two sources of evidence. The observed count matrix C from standard equilibrium simulations and the data from enhanced or biased sampling simulations E used to estimate the stationary vector. We can sample the posterior of reversible matrices with fixed stationary vector $\mathbb{P}(P|\pi, C)$ using the methods outlined in Section 4.4. An error model can be used to estimate the uncertainty in the stationary vector. Such a model describes the posterior of stationary vectors given the enhanced sampling data,

$$\mathbb{P}(\pi|E). \tag{4.36}$$

The posterior for reversible transition matrices under the combined evidence $\mathbb{P}(P|C, E)$ can be formally decomposed as

$$\mathbb{P}(P|C, E) = \int \mathrm{d}\pi \, \mathbb{P}(P|C, \pi, E) \mathbb{P}(\pi|C, E). \tag{4.37}$$

Assuming that the direct effect of the enhanced sampling information E is negligible in the posterior of transition matrices with given stationary vector,

$$\mathbb{P}(P|C, \pi, E) \approx \mathbb{P}(P|C, \pi) \tag{4.38}$$

and that the direct effect of observed transition counts C is unimportant compared to the enhanced sampling data used to obtain π,

$$\mathbb{P}(\pi|C, E) \approx \mathbb{P}(\pi|E) \tag{4.39}$$

we model the uncertainty encoded in (4.37) by inserting the two approximations (4.38), (4.39),

$$\mathbb{P}(P|C, E) \approx \int d\pi \, \mathbb{P}(P|C, \pi) \mathbb{P}(\pi|E). \qquad (4.40)$$

Approximate sampling from $\mathbb{P}(P|C, E)$ can now be achieved by drawing a random sample $\{\pi^{(1)}, \ldots, \pi^{(M)}\}$ distributed according to a given error model, $\pi^{(k)} \sim \mathbb{P}(\pi|E)$ and generating a sample of transition matrices $\{P_1^{(k)}, \ldots, P_N^{(k)}\}$ from the constrained posterior $P_i^{(k)} \sim \mathbb{P}(P|C, \pi^{(k)})$ for each of the $\pi^{(k)}$. The sample $\{P_1^{(1)}, \ldots, P_N^{(1)}, \ldots, P_1^{(M)}, \ldots, P_N^{(M)}\}$ will then be approximately distributed according to $\mathbb{P}(P|C, E)$.

ESTIMATION OF RARE EVENT KINETICS

In this chapter a new approach for the estimation of probabilities for rare transition events and related kinetic quantities is developed.

The detailed balance condition is used to estimate transition probabilities in situations in which only one direction of the process of interest has been observed *and* stationary probabilities for the process are available. Reversible Markov state models are used to combine enhanced sampling simulations and short relaxation trajectories generated by standard molecular dynamics simulations so that kinetics for rare event processes can be estimated without having to assume a rate model. It is demonstrated that the method can be used to obtain reliable estimates orders of magnitudes before a single rare event would have been observed on average. The material below has previously appeared in similar form in [113].

5.1 EFFICIENT ESTIMATION VIA DETAILED BALANCE

There are many methods that allow to efficiently estimate the stationary vector, even in situations in which a direct estimation from a finite observation of the Markov chain is infeasible due to the metastable nature of the system. In such situations it is often possible to alter the system dynamics in a controlled way such that the artificial dynamics equilibrates more rapidly than the original one [42, 59, 104, 110, 111, 122]. The desired stationary vector of the original dynamics can then be related to the stationary vector estimated from the altered process [5, 35, 57, 99, 106].

For given stationary vector (π_i) detailed balance enforces a linear dependence between the transition matrix element p_{ij} and the element p_{ji}. As an immediate consequence the relative standard deviation of both elements is equal for the posterior of reversible transition matrices (4.12) with fixed stationary vector

$$\frac{\sqrt{\mathbb{V}(p_{ji})}}{\mathbb{E}(p_{ji})} = \frac{\sqrt{\mathbb{V}(p_{ij})}}{\mathbb{E}(p_{ij})}. \tag{5.1}$$

In situations where we can efficiently sample one of the two transitions (the fast one) and we know the stationary vector we obtain a reliable estimate of the probability for the other transition (the slow one) by virtue of the detailed balance condition. In other words a potential speed-up can be achieved in situations for which $p_{ij} \ll p_{ji}$ *and* we have *a priori* knowledge about the stationary vector $\pi = (\pi_i)$.

We show that (5.1) can be used to estimate kinetics for metastable systems efficiently. The main idea is to combine *a priori* knowledge about the stationary vector with short simulations providing evidence for the probability of the fast transition. Enforcing detailed balance with respect to the given stationary vector enables us to accurately estimate the probability for the slow transition without having to observe it in simulations at all. Short simulations will usually start in a state of high energy (low probability) and relax towards a state of low energy (high probability).

We estimate kinetics and their uncertainties via the posterior of Markov models for three different settings:

1. MSMs are sampled using only long unbiased simulations (nonreversible posterior (4.7))

2. MSMs are sampled using short relaxation trajectories and *a priori* knowledge about the true stationary vector (reversible posterior (4.12) with fixed stationary vector)

3. MSMs are sampled using short relaxation trajectories and enhanced sampling simulations (posterior with uncertain stationary vector (4.40)).

For metastable systems we propose the following strategy for distributing initial conditions exploiting the information from the stationary vector. Once all metastable sets and all kinetic barriers separating the sets have been identified using some enhanced sampling protocol, short trajectories should be started on top of all barriers or in high-energy metastable states. The length of the short trajectories needs to be sufficient to relax towards the low-energy metastable states. The method described here can be used to combine these data into an estimate of the full rare event kinetics.

5.2   FINITE STATE SPACE MARKOV CHAIN

Consider a three-state Markov chain with the following transition matrix

$$P = \begin{pmatrix} 1 - 10^{-b} & 10^{-b} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 10^{-b} & 1 - 10^{-b} \end{pmatrix}. \tag{5.2}$$

The parameter $b > 0$ can be thought of as the height of an energy barrier between states one and three. The stationary vector for (5.2) is

$$\pi = (1 + 10^{-b})^{-1} \left(\frac{1}{2}, 10^{-b}, \frac{1}{2} \right)^T. \tag{5.3}$$

The vector π and the matrix P satisfy the detailed balance equation.

Any process starting in state one has an exponentially small probability of crossing over to state three. In fact a chain starting in state one can reach state three only via state two, but the probability to go from state one to state two is exponentially small in the barrier height b. The reversed process, going from state two to state one, occurs much faster. The same applies to state three and state two. The eigenvalues of the matrix in (5.2) are

$$\lambda_1 = 1 \quad \lambda_2 = 1 - 10^{-b} \quad \lambda_3 = -10^{-b}.$$

The slowest time-scale (largest relaxation time) in the system is given by,

$$t_2 = -\frac{1}{\log \lambda_2} \approx 10^b.$$

It is apparent from $t_2 \approx p_{12}^{-1}$ that estimates of t_2 and of p_{12} have similar relative standard deviation. The relative standard deviation ϵ for a matrix-element p_{ij} for sampling from the unconstrained posterior (4.7) is

$$\epsilon(p_{ij}) \approx \frac{1}{\sqrt{c_{ij}}}.$$

For $b = 4$ and a single chain of length $N \approx 7 \cdot 10^4$ steps starting in state one we can on average expect $c_{12} = 4$ resulting in a relative standard deviation of 50%. In order to decrease this down to 1% we need to run a chain of length $N \approx 50^2 \cdot 7 \cdot 10^4 \approx 2 \cdot 10^8$ steps. This is clearly an unsatisfactory situation and we would like to reduce the required simulation effort to reach a given error level.

In comparison, for an ensemble of M short chains of length L starting in state two one will on average observe a transition from state two to state one for every second chain, $c_{21} = M/2$, so that a relative standard deviation of 1% for p_{21} can already be achieved for $M \approx 10^4$, with $L \ll 10^b$. As a result, the total simulation effort can be reduced by orders of magnitude.

We do not have explicit expressions for the relative standard deviation of matrix elements p_{ij} when sampling from the posterior of reversible transition matrices with fixed stationary vector. It is however conceivable that the relative standard deviation of p_{21} can be reduced in the same way ($\epsilon(p_{21})$ scales as $1/\sqrt{c_{21}}$). The relation (5.1) guarantees that a small relative standard deviation for the probability p_{21} will also result in a small relative standard deviation for the small probability p_{12}.

In Figure 5.1 we compare the mean value and the relative standard deviation for a system with barrier parameter $b = 4$. For the long chain estimates are obtained via the nonreversible posterior (4.7). For the ensemble of short relaxation trajectories estimates are obtained via the reversible posterior (4.12) with fixed stationary vector. The long

chain was started in state one while the short trajectories were started in state two. It is apparent that using short relaxation trajectories together with *a priori* information about the stationary vector results in a three orders of magnitude smaller simulation effort for the estimation of t_2 with a prescribed error. In particular, estimation of t_2 can be conducted orders of magnitude before a direct simulation would on average encounter a single transition event.

This effect is even more pronounced when choosing $b = 9$ so that estimation via long trajectories sampling the rare event is hopeless. Using short relaxation trajectories in combination with the stationary vector one can accurately estimate t_2 with a total simulation effort of $N = 10^3$ steps, c.f. Figure 5.2. A good estimate can be obtained six orders of magnitude before on average a single rare event would have occurred.

5.3   DOUBLE-WELL POTENTIAL

In the next example the Markov state model is an approximation of a continuous-valued stochastic process - Brownian dynamics in a double-well potential. The potential is given by the following expression

$$V(x) = (x^2 - \sigma^2)^2 + \delta\sigma(\frac{1}{3}x^3 - \sigma^2 x). \tag{5.4}$$

The two minima of the potential at $\pm\sigma$ are separated by a maximum at $-\delta\sigma/4$, cf. Figure 5.3. The dynamics is governed by the following stochastic differential equation (SDE),

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}}dW_t. \tag{5.5}$$

The increments dW_t are the increments of the Wiener process. The inverse temperature $\beta = (k_B T)^{-1}$ controls the intensity of the stochastic fluctuations.

The SDE (5.5) defines a process where X_t sample from the canonical distribution,

$$\pi(x) = Z(\beta)^{-1}e^{-\beta V(x)}. \tag{5.6}$$

The temperature dependent constant $Z(\beta)$ is the partition function ensuring correct normalization, $\int dx\,\pi(x) = 1$. Spectral properties of this Markov process, such as the largest relaxation time scale can be computed from a spatial discretization of its associated transition kernel, cf. Section A.1.

Parameters for the potential in (5.4) are $\sigma = 2.2$ and $\delta = 0.1$. The SDE in (5.5) is integrated using an explicit Euler-scheme with time step $\Delta t = 10^{-3}$ and noise parameter $\beta = 0.4$.

Values computed from the spatial discretization of the transition kernel are used as reference values for comparison with estimates obtained from a Markov model. The transition kernel is discretized on
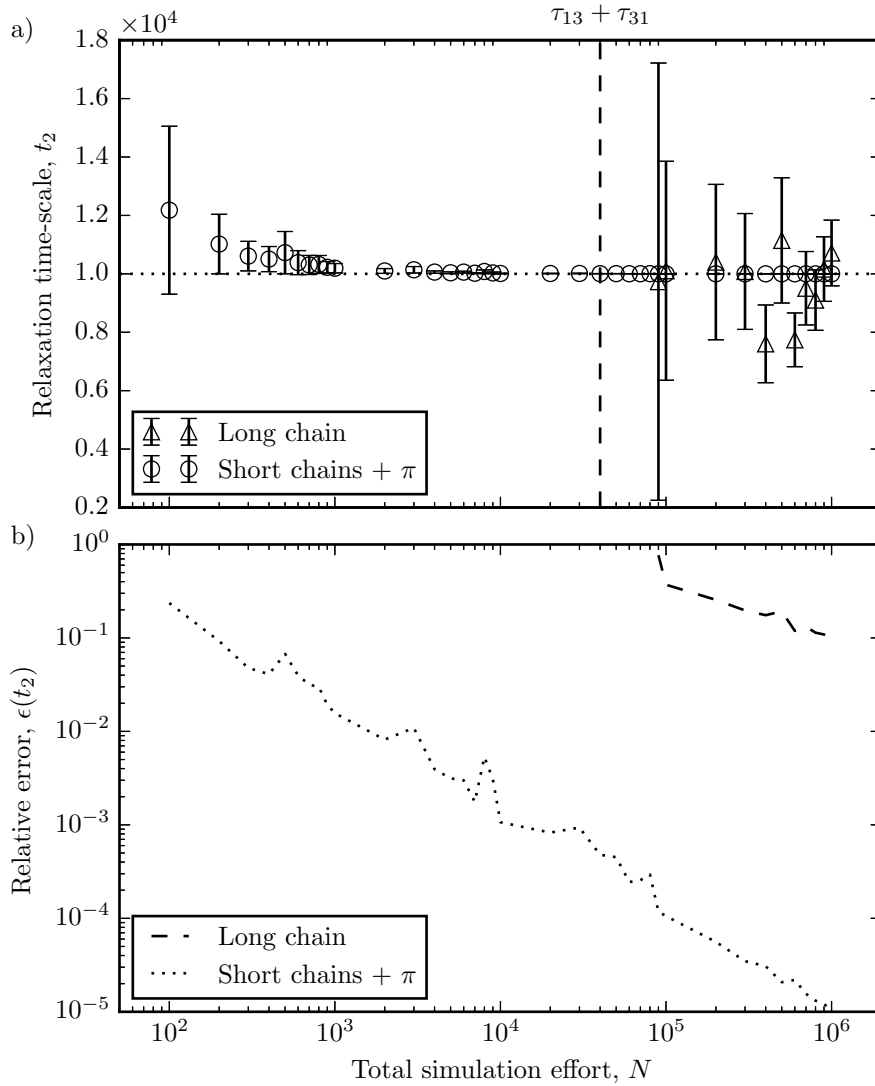
Figure 5.1: Convergence of mean value for metastable 3-state system with barrier parameter $b = 4$. The value t_2 is the largest relaxation time and N is the total simulation effort. We show mean value a) and relative standard deviation b) for a single long trajectory and for short relaxation trajectories with *a priori* knowledge about the stationary vector. The latter approach allows us to obtain a reliable estimate already before the average waiting time for a single rare event $\tau_{13} + \tau_{31}$ has elapsed. A comparison of the relative standard deviation indicates a three orders of magnitude speed-up when using short relaxation trajectories together with *a priori* knowledge about the stationary vector.
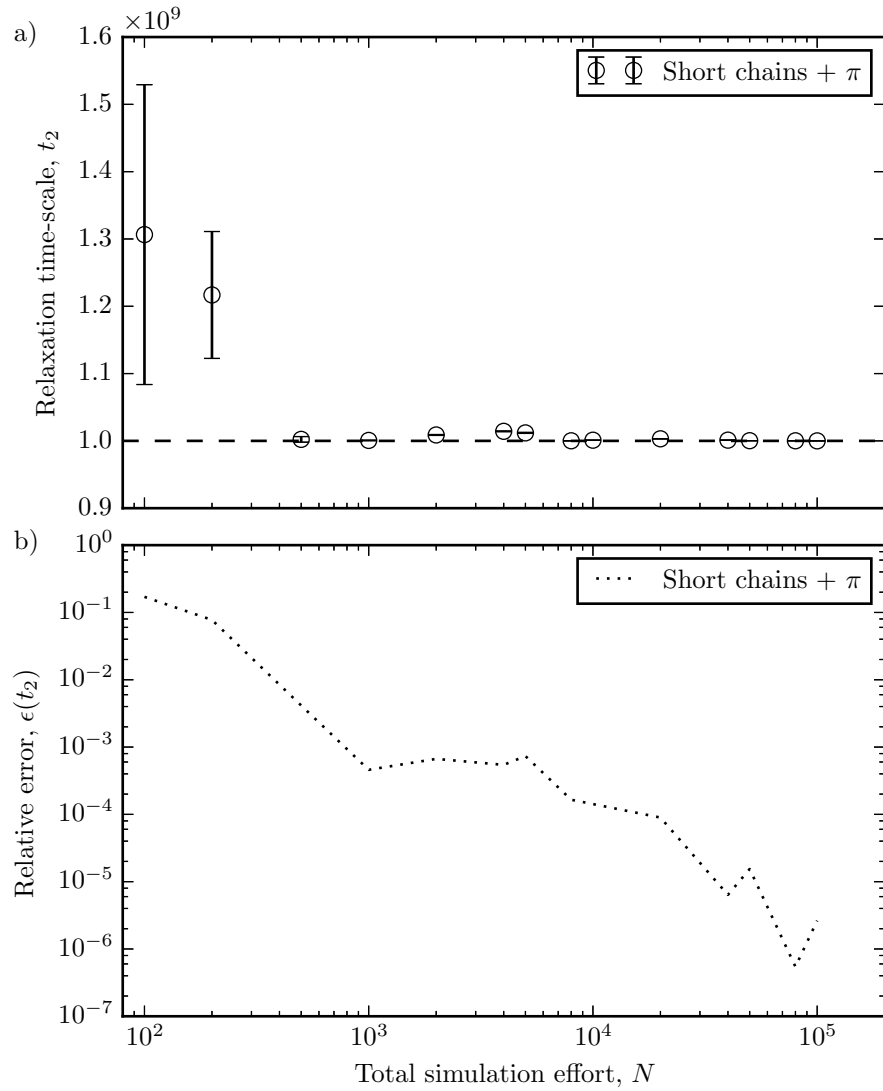
Figure 5.2: Convergence of mean value for metastable 3-state system with barrier parameter $b = 9$. The value t_2 is the largest relaxation time and N is the total simulation effort. We show mean value a) and relative standard deviation b) for short relaxation trajectories with *a priori* knowledge about the stationary vector. Estimation of t_2 via long trajectories is unfeasible. Convergence of the mean value indicates that a good estimate can be obtained six orders of magnitude before a single rare event would have occur-ed on average.
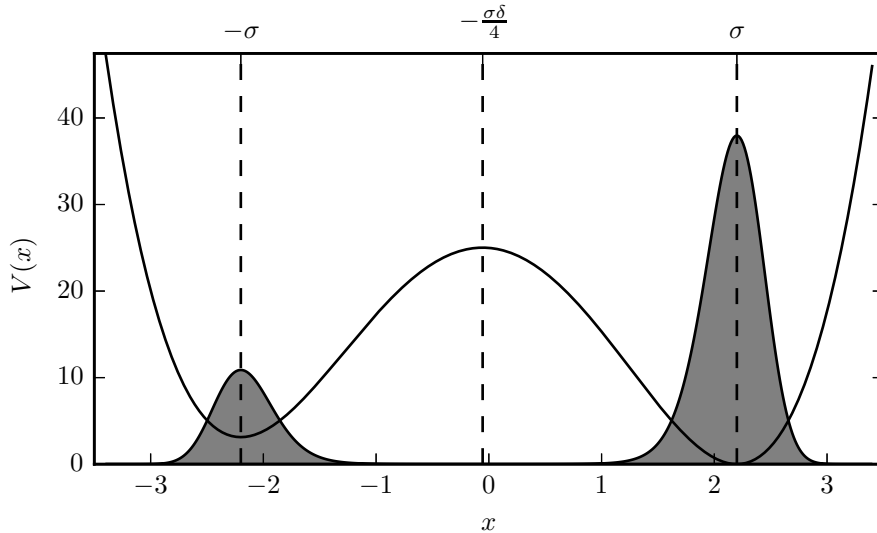
Figure 5.3: Double-well potential and corresponding equilibrium distribution for Brownian dynamics. The equilibrium distribution $pi(x)$ (shaded area) is scaled to fit the potential function $V(x)$. The equilibrium distribution is concentrated in the metastable regions around the two minima of the potential at $\pm\sigma$.

the interval $[-3.4, 3.4]$ using 400 equal sized bins. The matrix (p_{ij}) is assembled by evaluating the kernel at the midpoints of the bins. The largest relaxation time, $t_2 = 1.2 \cdot 10^6$, is computed from an eigenvalue decomposition of the assembled matrix. Mean first-passage times between sets $A = [\sigma - 0.2, \sigma + 0.2]$ and $B = [-\sigma - 0.2, -\sigma + 0.2]$ are computed, $\tau_{AB} = 5.3 \cdot 10^6$ and $\tau_{BA} = 1.6 \cdot 10^6$, see Section A.3 for details.

For Markov model building the interval $[-3.4, 3.4]$ is discretized into 100 equal sized bins, resulting in a MSM with 100 states. The lagtime $\tau = 10\, dt$ is determined from an implied time scale plot. Time scales are estimated from long trajectories ($N = 10^8$ steps).

We are interested in estimates of the largest relaxation time t_2. We compare estimates for long trajectories with estimates for short relaxation trajectories and umbrella sampling simulations. Long trajectories start in the metastable region, $x_0 = \sigma$, and short relaxation trajectories start on the barrier, $x_0 = -\delta\sigma/4$. Transition matrices are sampled according to (4.7) if only long trajectories are used and according to (4.40) if short relaxation trajectories are used and the stationary vector is estimated from umbrella simulation data.

The total simulation effort N is composed of the simulation effort spent on obtaining a count matrix from standard simulations, N_C, and the simulation effort spent on obtaining the equilibrium distribution from umbrella sampling simulations, N_π,

$$N = N_\pi + N_C. \tag{5.7}$$

The stationary vector is estimated from umbrella sampling simulations using the weighted histogram analysis method [35, 58]. Estimates are computed using 20 umbrella sampling simulations with $2.5 \cdot 10^4$ or $5 \cdot 10^5$ points per umbrella. To account for the uncertainty in the estimated stationary vector we use bootstrap resampling [28] of the generated data. The stationary vector is computed for each resampled data set to model the ensemble of stationary vectors compatible with the observed umbrella sampling data, i.e. the posterior distribution (4.36). The count matrix is computed either from $20 - 100$ long trajectories of length $10^6 \, dt$ or from $50 - 5000$ short relaxation trajectories of length $10^4 \, dt$.

In Figure 5.4 we show convergence of the posterior mean value of t_2 for four different data sets:

1. only long trajectories

2. short relaxation trajectories and a small amount of umbrella sampling data

3. short relaxation trajectories and a large amount of umbrella sampling data

4. an even splitting of the total effort into the effort for umbrella sampling simulations and the effort for short trajectories

It can be seen that for a fixed effort N_π the relative standard deviation cannot be reduced below a certain amount with increasing N_C. This is a result of the nonzero statistical error in the estimate of the equilibrium distribution for fixed N_π. The usual $N^{-\frac{1}{2}}$ dependence of the relative standard deviation can be recovered for the proposed splitting $N_\pi = N_C = N/2$. Figure 5.4 shows the favorable scaling coefficient of such an approach leading to a more than two orders of magnitude faster convergence of the estimated quantity compared to using standard simulations alone. Reliable estimates of the rare event kinetics can be obtained one order of magnitude simulation effort before the standard approach using long trajectories and no information about the equilibrium probabilities can be applied at all. The finite error for the estimate of the equilibrium distribution for $N_\pi = 5 \cdot 10^5 \, dt$ and $N_\pi = 10^7 \, dt$ results in a saturation of the error of t_2 which can be further decreased using a more precise estimate of the equilibrium distribution from additional enhanced sampling simulations.

5.4   ALANINE DIPEPTIDE

In this example the Markov state model is an approximation to a projection of a high-dimensional continuous-valued stochastic process onto a set of coordinates of low dimension. The ϕ and ψ dihedral angles have been identified as the two relevant coordinates for the slowest kinetic
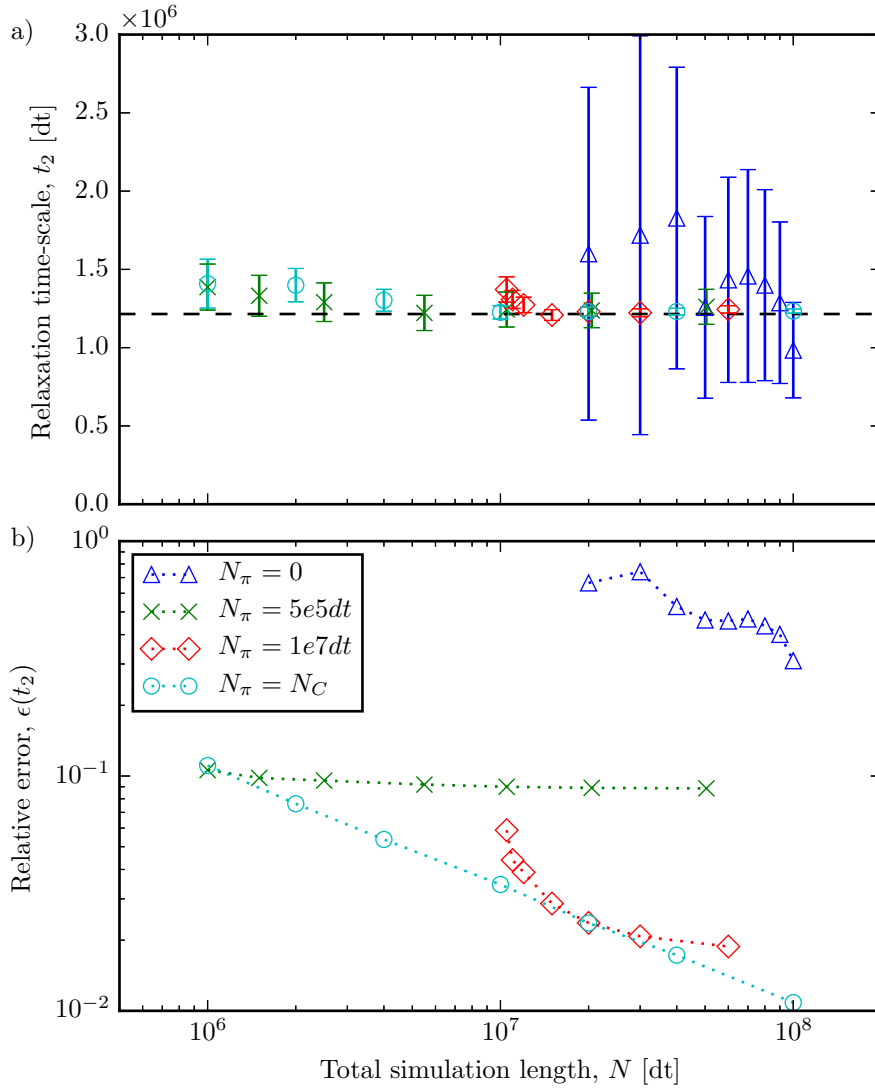
Figure 5.4: Convergence of mean value for Brownian dynamics in double-well potential. The value t_2 is the largest relaxation time and N is the total simulation effort. We show mean value a) and relative standard deviation b) for a single long trajectory (blue) and for different combinations of short relaxation trajectories with enhanced sampling simulations (green, red, light-blue). The latter approach allows to obtain a reliable estimate using almost one order of magnitude less simulation effort than the average waiting time for a single rare event $\tau_{AB} + \tau_{BA} = 7 \cdot 10^6 dt$. A comparison of the relative standard deviation shows a more than two orders of magnitude speed-up when combining short relaxation trajectories and enhanced sampling simulations.
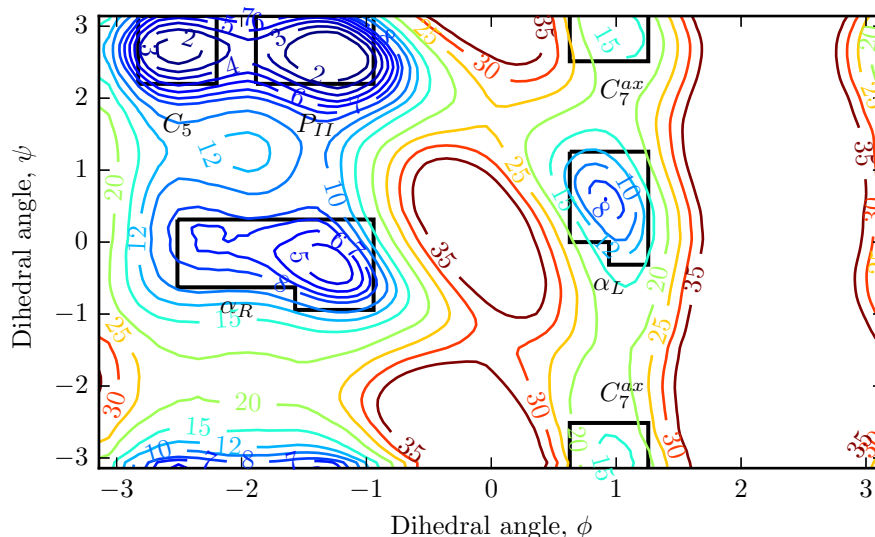
Figure 5.5: Free-energy profile for alanine dipeptide ϕ, ψ dihedral angle data. Energies values are in kJ/mol. The average thermal energy $k_B T$ at $300K$ is $2.493 kJ/mol$. One can identify five metastable sets on the dihedral angle torus, here indicated by black lines. There are three low energy (high probability) sets C_5, P_{II} and α_R with $\phi < 0$ and two high energy (low probability) sets α_R and C_7^{ax} with $\phi > 0$.

processes of the system in equilibrium [1]. The potential of mean force for the two dihedral angles is shown in Figure 5.5. As an example for a rare event quantity in a molecular system we use the mean first-passage time for the C_5 to α_L transition in the alanine dipeptide molecule.

Alanine dipeptide has been the long-serving laboratory rat of molecular dynamics [1, 18, 26, 67, 109]. One can identify five metastable regions in the free-energy landscape. The C_5 and P_{II} regions correspond to dihedral angles found in a beta-sheet conformation, the α_R and α_L regions correspond to a right, respectively left-handed α-helix conformation. Reference values for the mean-first passage times between all pairs of sets have been computed from the maximum likelihood estimator of (2.45) using a total of $10\mu s$ of simulation data. Reference values can be found in Table 5.1. For details of the computation of mean first-passage times see Section A.3.

All molecular dynamics simulations were carried out on graphics processing units (GPU) using the OpenMM simulation package [27]. We use the *amber99sb-ildn* forcefield [62] and the *tip3p* water model [50]. The peptide was simulated in a cubic box of $2.7nm$ length including 652 solvent molecules. Langevin equations were integrated at $T = 300K$ using a time-step dt of $2fs$. The potential used for umbrella sampling simulations was $V_i(\phi) = k[1 + \cos(\phi - \phi_i - \pi)]$ with $k = 200 kJ/mol$. Umbrellas were placed at a spacing of $\phi_i - \phi_{i+1} = 9°$.

τ_{AB}/ns	C_5	P_{II}	α_R	α_L	C_7^{ax}
C_5	0	0.021	0.253	43.456	60.220
P_{II}	0.041	0	0.255	43.449	60.213
α_R	0.142	0.125	0	43.549	60.312
α_L	1.553	1.527	1.744	0	17.757
C_7^{ax}	1.559	1.533	1.745	1.221	0

Table 5.1: Mean first-passage times for alanine dipeptide. Mean first-passage times (MFPT) are computed for transitions between metastable regions. The MFPTs have been estimated from $10\mu s$ of simulation data using a Markov state model.

5.4.1 *Analysis in ϕ and ψ dihedral angle space*

We show the convergence of the largest relaxation timescale and validate the MSM constructed at a lagtime of $\tau = 6ps$ via a Chapman-Kolmogorov test in Figure C.1. Convergence of the largest relaxation time indicates that the slow eigenfunctions of the associated dynamical operator are well approximated by the discrete MSM. The Chapman-Kolmogorov-test explicitly checks the Markov assumption comparing self transition probabilities computed from the MSM, parameterized at lagtime τ, with direct estimates from the data at larger lagtimes, $n\tau$. A thorough discussion of MSM validation can be found in [84].

In Figure 5.6 we show the estimate of the mean first-passage time τ_{AB} between the C_5 and the α_L region together with the corresponding relative standard deviation $\epsilon(\tau_{AB})$ for different values of the total simulation effort N. The simulation setup is similar to the one described for the double-well potential in the previous section. Instead of starting short trajectories directly on the barrier we start them from the metastable α_L region. Figure 5.6 shows that combining umbrella sampling data and short trajectories relaxing from a metastable region with low probability (high free-energy) towards a metastable state with high probability (low free-energy) is able to estimate the reference value, $\tau_{AB} = 43ns$ for the C_5 to α_L transition with a total simulation effort of $70ns$ if short 'downhill' trajectories are used in combination with umbrella sampling data. Utilizing information about the stationary vector in combination with short simulations that do not have to sample the rare event is able to achieve a relative standard deviation with almost an order of magnitude less simulation effort compared to an ensemble of long trajectories. The observed 8-fold speed-up is in good agreement with the expected speed-up given by

$$\frac{\tau_{AB}}{L}$$

with $\tau_{AB} = 43ns$ the MFPT for the slow "up-hill" transition from C_5 to α_L, and $L = 5ns$ the length of individual short trajectories.
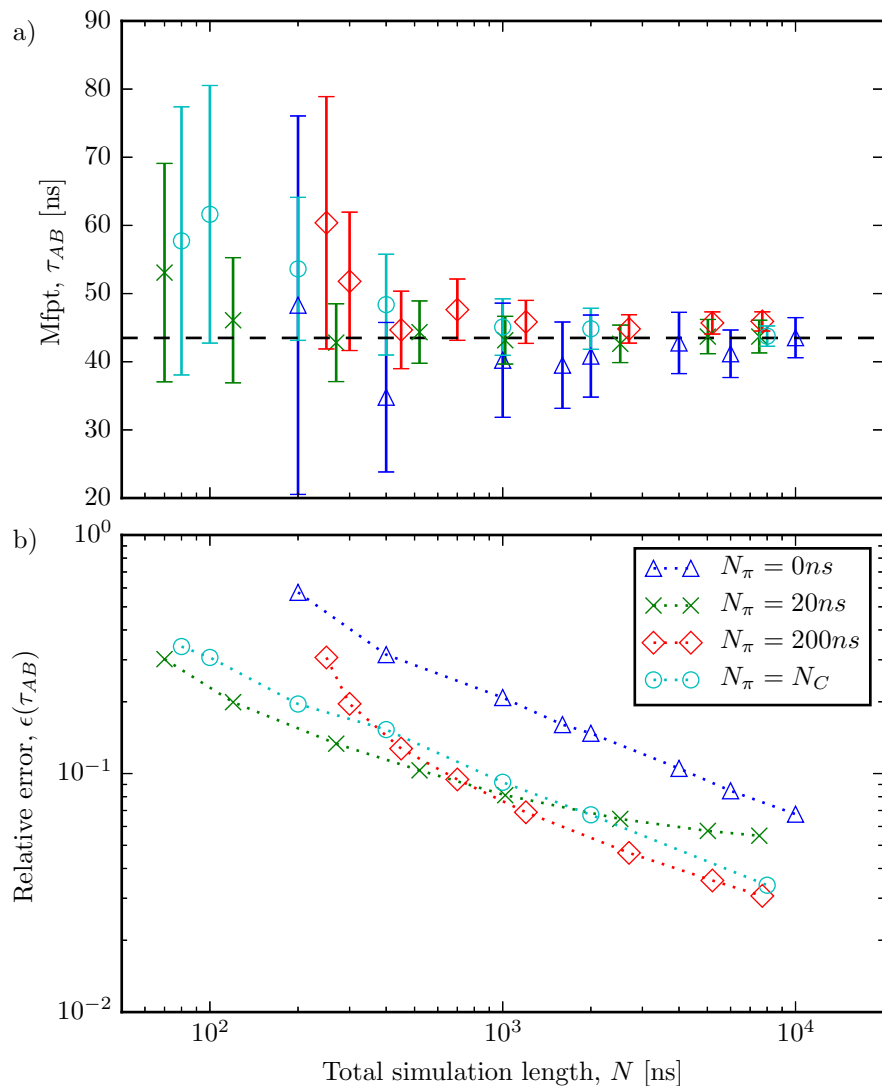
Figure 5.6: Convergence of mean value for alanine dipeptide ϕ, ψ dihedral angle data. The value τ_{AB} is the mean first-passage time of the rare C_5 to α_L transition and N is the total simulation effort. We show mean value a) and relative standard deviation for a small number of long chains starting in the C_5 region (blue), an ensemble of short relaxation trajectories starting in the α_L region combined with different amounts of umbrella sampling simulations (green, red, light-blue). A reliable estimate of $\tau_{AB} = 43ns$ can be obtained already for a total simulation effort of $N = 70ns$ when short relaxation trajectories are used in combination with umbrella sampling simulations. A comparison of the relative standard deviation shows a speed-up of almost one order of magnitude compared to using only long trajectories.
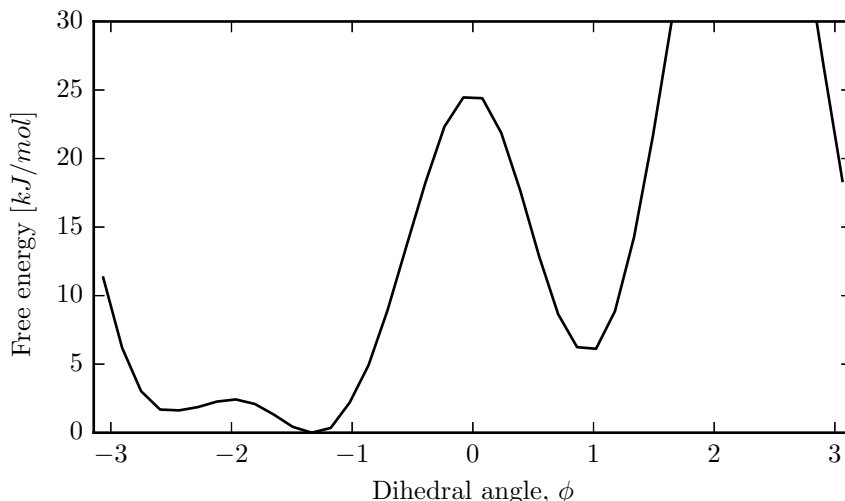
Figure 5.7: Free-energy profile for alanine dipeptide ϕ dihedral angle data. One can identify three metastable sets. Two low energy (high probability sets) with $\phi < 0$ and a single high energy (low probability) set with $\phi > 0$.

The present approach of estimating rare event kinetics is more powerful than traditional rate theories because quantities that can be estimated can be much more complex than only rates. As a reversible Markov model is estimated, full mechanisms, such as the ensemble of transition pathways from one state to another state can be computed.

5.4.2 *Analysis in the ϕ-coordinate alone*

The presented method can also work if only information about the slowest degree of freedom is used. In Figure 5.7 we show the free-energy profile for the ϕ dihedral angle. An energetic barrier clearly separates the low free-energy region, $\phi < 0$ from the high free-energy region, $\phi > 0$. Crossing events from $\phi < 0$ to $\phi > 0$ are rare leading to a sampling problem if kinetic quantities associated with barrier-crossings need to be estimated. Again we show convergence of the largest relaxation timescale, t_2 and a Chapman-Kolmogorov test for a MSM estimated at lagtime $\tau = 15ps$, Figure C.2.

In Figure 5.8 we show that the correct mean first passage time for the C_5 to α_L transition can also be recovered from the MSM of the ϕ angle alone. This demonstrates that the presented method is robust with respect to the choice of microstates. Choosing a slightly larger lagtime $\tau = 15ps$ for the ϕ MSM allows us to recover the correct mean first-passage time despite the fact that information about the ψ dihedral angle is completely neglected. The MSM for ϕ dihedral angle is still a good approximation to slowest kinetics if the discretization and the lagtime are suitably matched. A thorough discussion of approximation errors for MSMs can be found in [84, 91].
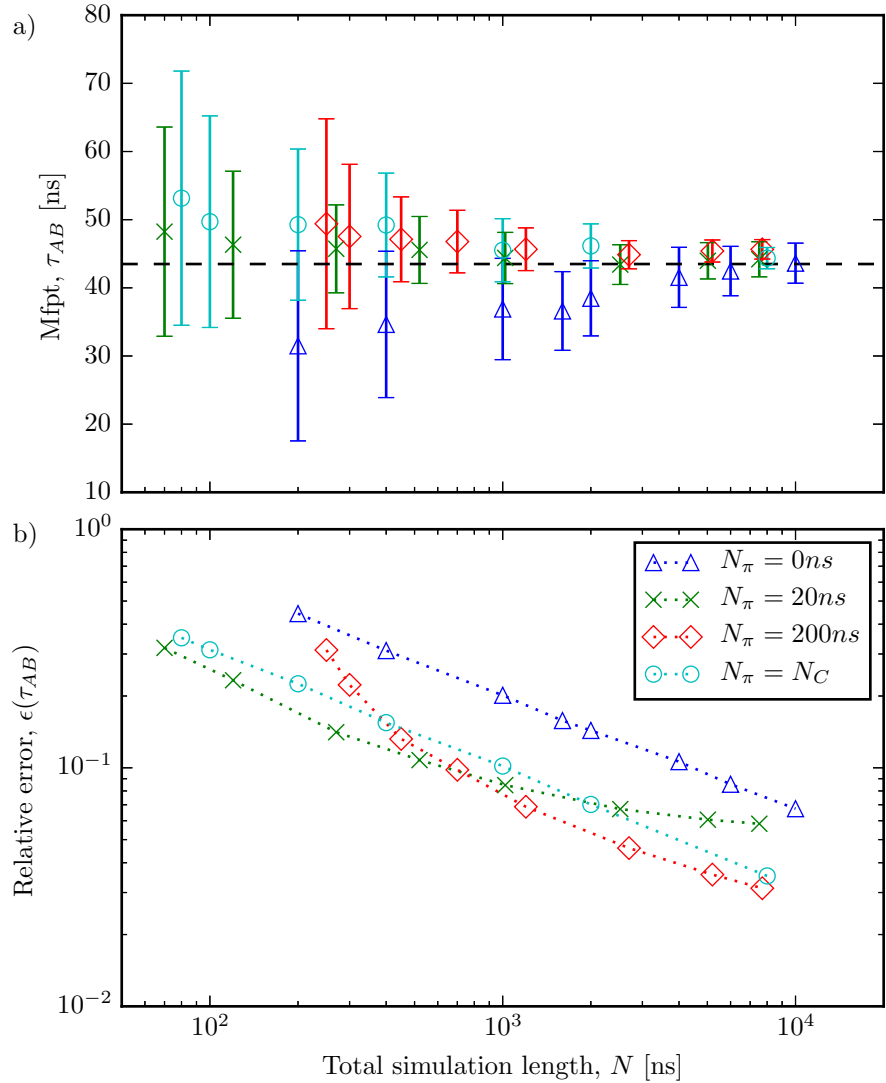
Figure 5.8: Convergence of mean value for alanine dipeptide ϕ dihedral angle data. The value τ_{AB} is the mean first-passage time of the rare transition from the low free-energy region $A = \{\phi| -162° < \phi < -54°\}$ to the high free energy region $B = \{\phi|36° < \phi < 72°\}$ and N is the total simulation effort. We show mean value a) and relative standard deviation for a small number of long chains starting in the A region (blue), an ensemble of short relaxation trajectories starting in the B region combined with different amounts of umbrella sampling simulations (green, red, light-blue). A comparison of the relative standard deviation shows a speed-up of almost one order of magnitude compared to using only long trajectories. The correct value, $\tau_{AB} = 43ns$, for the C_5 to α_L transition can be obtained even if no information about the ψ dihedral angle is used in the construction of the MSM.
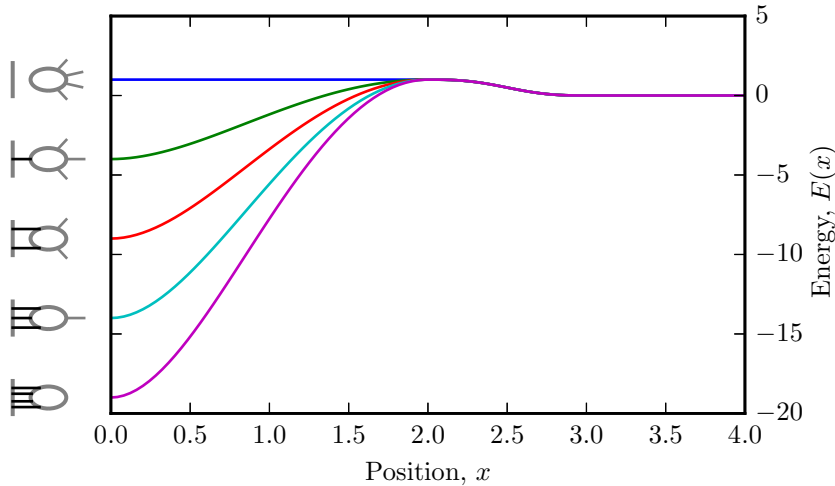
Figure 5.9: Energy landscape for vesicle model. We show the vesicle binding energy for different attachment modes, $m = 0, 1, \ldots, 4$.

5.5   VESICLE MODEL

As a final example we consider the diffusive motion of a colloid that can reversibly attach to a surface via $m = 0, \ldots, M$ tethers. A biological example of such a system is a neuronal vesicle that can attach to a plasma membrane by SNARE protein complexes. The diffusion in the solvent is free but the attachment of tethers restricts the location of the vesicle to a vicinity of the membrane. The restriction is stronger the more tethers are attached. Attachment of the vesicle to the membrane is a fast process, but the dissociation from the membrane is an extremely rare event. We show that the mean first passage time for dissociation can be reliably estimated despite the fact that a non-Markovian coordinate, the membrane-vesicle distance, is used.

Figure 5.9 shows the energy for the different vesicle attachment modes. For $m > 0$ attachment of the vesicle to the membrane is governed by a harmonic potential close to the membrane. For $x > 2$ all attachment modes are energetically equal corresponding to a breaking of the m tethers once the distance between the vesicle and the membrane exceeds a certain threshold. The association of the vesicle has to overcome a small energetic barrier, modeling a weak repulsion of the untethered vesicle.

The state of the vesicle is given by the pair (x, m) where x is the vesicle membrane distance and m denotes the number of tethers attached. A discretization of the vesicle membrane distance with $0 = x_1 < \cdots < x_d = 4$ allows us to describe the vesicle dynamics by a Markov chain on a finite state space with $(M + 1)d$ microstates. The stationary vector of the chain is

$$\pi = (\pi^{(0)}(x_1), \ldots, \pi^{(M)}(x_d)) \tag{5.8}$$

with entries given in terms of the usual Gibbs/Boltzmann distribution,

$$\pi^{(m)}(x_i) \propto e^{-E^{(m)}(x_i)}. \tag{5.9}$$

$E^{(m)}(x)$ is the energy of a vesicle at x with m tethers attached, cf. Figure 5.9. The energy of the different attachment modes is given by

$$E^{(m)}(x) = \begin{cases} 1 + m(-5 + 5x^2 - 2.5x^3 + 0.3125x^4) & 0 \leq x < 2 \\ 1 + 8(x-2)^2 - 8(x-2)^3 & 2 \leq x < 2.5 \\ 0.5 - 8(x-2.5)^2 + 8(x-2.5)^3 & 2.5 \leq x < 3 \\ 0 & 3 \leq x < 4 \end{cases} \tag{5.10}$$

The transition matrix $P = (p_{ij})$ for the vesicle dynamics is now constructed as follows. We encode random walk probabilities in a proposal matrix $Q = (q_{ij})$. The particle moves from x_i to x_{i-1} or x_{i+1} with probability $1/3$, if the particle remains at its current position x_i it can attach, $m \to m+1$, or detach, $m \to m-1$, a tether with probability $1/3$ so that the overall proposal-probability for attachment or detachment is $1/9$. To account for the energetic differences of the microstates we use the Metropolis-Hastings acceptance criterion to modulate the proposal probabilities and obtain the desired transition probabilities via,

$$p_{ij} = \min\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\} \quad i \neq j. \tag{5.11}$$

Correct normalization is ensured by setting $p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. As a result of (5.11) the constructed transition matrix P automatically fulfills the detailed balance condition (3.1) with respect to the desired equilibrium distribution.

The mean first-passage time for the dissociation of the vesicle is $\tau_{AB} = 8.56 \cdot 10^9$, the mean first-passage time for association, $\tau_{BA} = 1.59 \cdot 10^3$, is orders of magnitude smaller. The mean first-passage time for dissociation of a vesicle with the maximum number of tethers attached is $\tau_{AB} = 3.83 \cdot 10^{10}$ so that the system dynamics cannot be described in terms of the subspace with $m = 4$ tethers. This indicates that the dissociation kinetics is effectively non-Markovian along the x-coordinate.

The dissociation time τ_{AB} can reliably be estimated even if no information about the mode of attachment is available. If only information about the position of the vesicle is available then the state-space of the $(M+1)d$ distinct microstates is coarse-grained into d distinct sets each containing $(M+1)$ microstates corresponding to the $M+1$ possible tethering modes at position x. The coarse grained equilibrium distribution $\tilde{\pi}$ is obtained by summing the full equilibrium distribution π over all possible tethering modes. If short association trajectories starting in the region $x > 2$ are combined with the coarse-grained equilibrium

distribution the dissociation time can again be estimated orders of magnitude before a single dissociation event would on average be observed despite the fact that the MSM is built on a coordinate that is inherently non-Markovian. In Figure 5.10 we show mean and standard error for the MFPT of vesicle dissociation for a MSM build at a lagtime of $\tau = 60$ with $d = 40$ microstates.

In Figure C.3 we again show convergence of the largest relaxation time and the Chapman-Kolmogorov test for an MSM constructed at lagtime $\tau = 60$. The MSM is estimated solely from short association trajectories starting in the high energy region using the coarse grained equilibrium distribution $\tilde{\pi}$ to obtain a reversible maximum likelihood transition matrix from (3.26). The total simulation effort, $N = 2 \cdot 10^7$, used to obtain the MSM and perform the validation is again orders of magnitude smaller than the expected dissociation time.
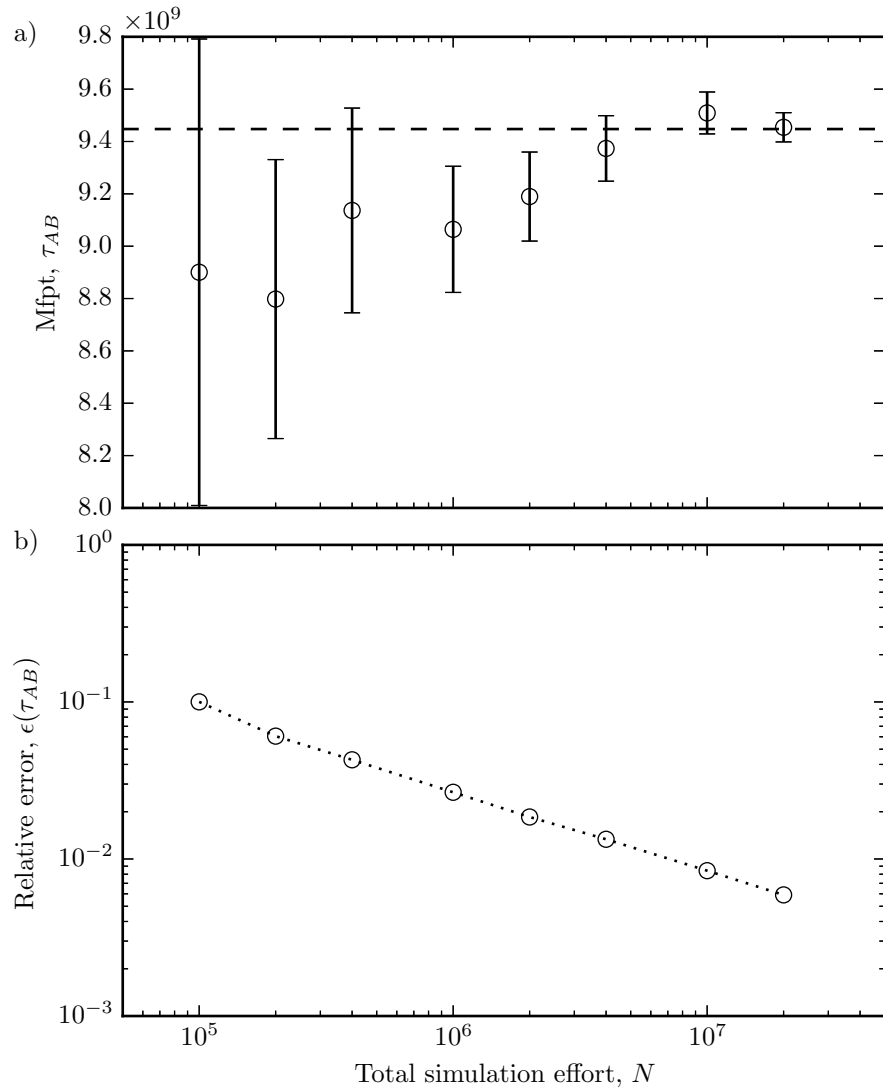
Figure 5.10: Convergence of mean value for vesicle model. The value τ_{AB} is the mean first-passage time for the vesicle dissociation. Estimation is performed on a coarse-grained non-Markovian state space. We show mean value a) and relative standard deviation b) for an ensemble of short relaxation trajectories together with *a priori* knowledge about the coarse-grained stationary vector. The dissociation time can be estimated orders of magnitude before a single dissociation event would have been observed.

CONCLUSION

Reversible Markov state models are the discrete space, finite data counterpart to the transfer operator for conformation dynamics. Reversible Markov state models arise naturally from molecular dynamics implementations which are often constructed in such a way that they obey reversibility or at least a generalization thereof [60, 116, 118]. Reversibility of the dynamics results in a detailed balance equation for the transition probabilities. In order to have reversibility in a Markov model, detailed balance has to be enforced during the estimation.

Maximum likelihood estimators for transition probabilities which obey the detailed balance equation can be constructed. Reversible maximum likelihood estimation of transition probabilities results in a large nonlinear optimization problem. The problem has to be solved numerically which can be difficult; the problem is nonconvex and the number of unknowns in the problem scales quadratically with the number of states of the Markov model. A duality argument from [123] can be used to transform the optimization problem into a constrained saddle-point problem (convex-concave program) with a number of unknowns scaling linearly with the number of states. An efficient algorithm for the solution of the convex-concave program based on a primal-dual interior-point method from [86] is developed. The method is superlinearly convergent, robust to the input data, and can be used to estimate reversible Markov models with a large number of states. The algorithm can also be used to estimate reversible Markov state models if *a priori* knowledge about the stationary probabilities is available.

Methods for uncertainty quantification of reversible Markov state models are developed. A prior distribution for transition matrices fulfilling the detailed balance equation is constructed and an efficient Monte Carlo algorithm is developed to sample reversible transition matrices. The constructed prior encodes the following *a priori* assumption: All transitions for which neither the forward nor the backward transition has been observed in the data are forbidden (probability of zero). The prior enforces the same sparsity structure for the sampled matrices as the symmetrized matrix of transition counts. This prior leads to error bars which nicely envelope reference estimates. Posterior sampling of reversible transition matrices is achieved by a Gibbs sampling algorithm for which the proposal distribution of each Gibbs sampling step is adapted to match the conditional posterior distribution. The algorithm can be used to compute uncertainties of Markov models with a large number of states for simulation data containing rare events.

Reversible Markov state models can be used to estimate probabilities for rare transitions. If detailed balance is enforced estimates can

be computed efficiently in situations in which only one direction of the process has been observed *and* stationary probabilities for the states of the MSM are available. Stationary probabilities can be estimated from enhanced sampling simulations which do not suffer from the sampling problem of standard MD. A combination of enhanced sampling simulations and short relaxation trajectories can result in orders of magnitude speed up for estimation of transition probabilities. Kinetics for rare event processes can be estimated without having to assume a rate model whenever the kinetic properties of interest can be computed from a Markov model discretization of the system. The approach can be used to estimate rates and passage times, but also complex quantities such as committor functions and transition path ensembles.

Implementation of all algorithms described here are available via the python package msmtools [112] and the python Markov-modeling software PyEMMA [92]

Previously the maximum likelihood estimation problem for reversible transition matrices was solved by a self-consistent iteration [12, 84, 123]. The iteration can take a long time to converge and the convergence behavior can be sensitive to the data that is used. The developed algorithm achieves a similar good scaling (quadratic scaling of the running time in the number of states) as the self-consistent iteration, but it can be much more efficient.

The discrete transition matrix reweighting analysis method (dTRAM) is a general framework to combine enhanced sampling simulations and equilibrium MD simulations into an estimate of a reversible transition matrix at the interesting unbiased condition [123, 124]. A numerical solution of the dTRAM problem via the self-consistent iteration outlined in [123] can take a long time to converge. In this thesis the algorithm for the reversible MLE problem is extended so that it can also solve the dTRAM problem. It is demonstrated that the proposed algorithm is able to solve the dTRAM problem orders of magnitudes faster than the self-consistent iteration.

Previous methods for sampling of reversible transition matrices were outlined in [63, 65, 73]. In [63, 73] the problem of prior choice for simulations containing rare events was not discussed. As shown in the thesis the choice of prior distribution is important to ensure that error bars computed via posterior sampling envelope the true value for simulation data containing rare events. In [65] a penalty prior for transition matrix sampling was proposed to address this problem. The prior results in samples which respect the sparsity structure of the nonreversible maximum likelihood estimate but not the sparsity structure of the reversible maximum likelihood estimate. A prior enforcing the reversible sparsity structure is also proposed. In contrast to the prior constructed here, this prior does not arise from a simple algebraic expression. Reversible sampling methods in [63, 73] use uniform proposal densities resulting in samplers with large autocorrelation times. The method in

[4] use a conjugate prior on the set of reversible transition matrices but an appropriate choice of prior (hyper-)parameters for simulation data containing rare events is not discussed. A method to sample reversible transition matrices with fixed stationary vector was also outlined [73]. It is demonstrated that the algorithm outlined here achieves much smaller autocorrelation times for the sampling of relevant kinetic observables.

The combination of enhanced sampling simulations and short relaxation trajectories to estimate probabilities for rare transitions is entirely new and has been presented in [113] for the first time. Previous methods [7, 31, 56, 108] assumed a specific rate model in order to reconstruct the kinetics from the free energy profile.

Reversible Markov state models can only describe fluctuations of a system in (thermal) equilibrium. If time-reversibility of the dynamics is destroyed, e.g. because a time-dependent external force driving the system out of equilibrium is applied, then one cannot describe the dynamics by a reversible model.

Successful construction of a Markov state model requires that the resolved coordinates are slow compared to the non-resolved coordinates (approximate Markovianity at chosen lagtime). Furthermore the chosen discretization needs to resolve the slow processes (discretization error). Slow coordinates and discretization have to be chosen either based on prior knowledge about the processes of interest or they need to be inferred on the basis of simulation data. In most cases slow processes are not known beforehand and simulations are carried out in order to uncover them. Furthermore, molecular simulations often contain very few observations of the relevant process which makes identification of a good discretization difficult. The thesis does not discuss this important aspect of Markov model building. Many approaches have been developed to tackle the problem. One example is the time-lagged independent component analysis method (TICA) [66, 83, 97]. TICA estimates a linear transformation that can be used to extract the linear subspace of slow coordinates from a given molecular dynamics trajectory.

The Markov state model approach relies on the ability to estimate the transition matrix and to use it for successive computations of quantities of interest. It is necessary that the slow conformational processes can be described by a small number of coordinates since the number of transition matrix elements that need to be estimated and stored scales exponentially with the number of coordinates. Hierarchical methods [80] can circumvent the curse of dimensionality because they approximate the transition matrix using a nested hierarchy of lower dimensional objects which can be stored and computed with computational effort that grows only polynomially in the number of coordinates. The method in [80] is based on a variational principle for the identification of slow conformational processes [74, 79]. Variational methods offer a greater flexibility in the choice of ansatz functions for the discretization

of the transfer operator than MSMs which require the use of characteristic functions. Relevant objects in a variational approach also need to be estimated on the basis of simulated trajectories so that they can also suffer from the sampling problem. In contrast to Markov state models, the analysis of statistical errors for variational methods has not been developed yet.

Fully automated construction of Markov models is still challenging. The discretization and lagtime for the Markov model have to be selected by hand. An automatic selection of the discretization and lagtime would require to simultaneously estimate the discretization error *and* the statistical error for given simulation data.

The formulation of the reversible posterior assumes independent counts, i.e. Markovianity at the chosen lagtime. This is not optimal since the required subsampling of the input trajectories results in lots of discarded data which could be used in the estimation procedure. A possible approach to circumvent this problem by computing effective counts from a time series using statistical inefficiencies is outlined in [71].

The proposed method for the efficient estimation of probabilities for rare transitions requires short relaxation trajectories that connect metastable conformations (in one direction) either by starting on an energetic barrier or by using an imbalance in the free energy of the two metastable sets. For the method, additional knowledge about the system is needed. The method requires *a priori* information about the stationary probabilities to infer the location of barriers and metastable sets. In contrast to milestoning or transition state sampling the method does not require to resolve the transition path very accurately.

The treatment of enhanced sampling simulations and equilibrium trajectories is decoupled. Information from transitions in the enhanced sampling simulations is neglected in the estimation of transitions in the unbiased ensemble and information from the unbiased simulation is neglected in the estimation of the stationary probabilities. In contrast to dTRAM, different lagtimes can be used for the estimation of stationary probabilities and transition probabilities.

In an adaptive sampling setup, in which enhanced sampling and equilibrium MD trajectories are combined, the developed methods can be used to decide if additional enhanced sampling simulations or additional equilibrium simulations should be performed to optimally distribute the computational effort.

A potential future application of the method is the estimation of the binding and unbinding kinetics of drug molecules to target proteins. The unbinding kinetics can be critical for drug efficacy [117] but it is usually very difficult to estimate. Often, the binding of the drug occurs relatively fast [14] but the dissociation may be orders of magnitude slower. The proposed method requires only trajectories sampling the fast binding event and information about the stationary probabilities of the process to estimate the probability for the slow unbinding event.

MSM - ANALYSIS

A.1 THE TRANSITION KERNEL FOR THE EULER-METHOD

The solution of (5.5) with initial position $X_0 = x_0$ on $[0, T]$ is usually carried out by choosing a regular discretization of the time interval

$$0 = t_0 < t_1 < \cdots < t_N = T.$$

with $\Delta t = t_k - t_{k-1}$ for all $k = 1, \ldots, N$. The evolution of the stochastic process is then approximated by the following time-stepping scheme

$$X_{t+\Delta t} = X_t - \nabla V(X_t)\Delta t + \sqrt{2\beta^{-1}}\eta. \tag{A.1}$$

with $X_0 = x_0$ and η being a $\mathcal{N}(0, \Delta t)$ distributed random variable. The time-stepping scheme (A.1) is known as *Euler method* or *Euler-Maruyama method*, [55].

For this simple time-stepping scheme the transition kernel of the resulting Markov chain is given by

$$p_{\Delta t}(x, y) = \frac{1}{\sqrt{2\pi\Delta t 2/\beta}} \exp\left(-\frac{(y - x + \nabla V(x)\Delta t)^2}{2(\sqrt{\Delta t}\sqrt{2/\beta})^2}\right), \tag{A.2}$$

with $x = X_t$ and $y = X_{t+\Delta t}$. The kernel $p_{\Delta t}(x, y)$ is a Gaussian distribution with mean $\mu = x - \nabla V(x)\Delta t$ and variance $\sigma^2 = 2\Delta t/\beta$.

The transition probability $P_{\Delta t}(B|A)$ between two sets A, B can be computed from

$$P_{\Delta t}(B|A) = \frac{\int_A \mathrm{d}x\, \pi(x) \int_B \mathrm{d}y\, p_{\Delta t}(x, y)}{\int_A \mathrm{d}x\, \pi(x)}. \tag{A.3}$$

Choosing a L such that $p_{\Delta t}(x, y)$ is effectively zero outside of $[-L, L]$ we pick a spatial discretization

$$-L = x_0 < x_1 < \ldots < x_N = L \tag{A.4}$$

with a regular spacing $\Delta x = x_k - x_{k-1}$ for $k = 1, \ldots, N$ such that $p_{\Delta t}(x, y)$ and $\pi(x)$ are approximately constant on sub-intervals $S_i = (x_k, x_{k+1}]$. In this case we have

$$\int_{x_i}^{x_{i+1}} \mathrm{d}x\, \mu(x) \approx \mu(x_k)\Delta x$$

and

$$\int_{x_i}^{x_{i+1}} \mathrm{d}x\, \mu(x) \int_{x_j}^{x_{j+1}} \mathrm{d}y\, p(x, y) \approx \mu(x_i)p(x_i, x_j)(\Delta x)^2.$$

We can approximate the matrix elements $p_{ij} = P(S_j|S_i)$ as

$$p_{ij} \approx p(x_i, x_j)\Delta x.$$

and compute spectral properties from the matrix (p_{ij}) using standard eigenvalue solvers.

A.2   EIGENVALUES AND RELAXATION TIMESCALES

Let P be a transition matrix for a Markov state model (MSM) at lagtime τ satisfying conditions (M1)-(M4) ,then P has the following spectral decomposition

$$P = \sum_{i=1}^{d} \lambda_i \phi_i \psi_i^T \tag{A.5}$$

with real eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d > -1$ and corresponding right and left eigenvectors ϕ_i and ψ_i.

The left eigenvalue corresponding to the eigenvalue $\lambda_1 = 1$ is the unique stationary vector $\psi_1 = \pi$ and the corresponding right eigenvalue is a constant $\phi_1 = (1, \ldots, 1)^T$. Since P satisfies the detailed balance equation the left eigenvectors ψ_i can be related to the right eigenvectors via $\psi_i = \pi \phi_i$ for all $i = 1, \ldots, d$.

For a metastable system with transfer operator P_t satisfying condition (C2) one can assume that for the corresponding transition matrix P there are m dominant eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_m > r > 0$ and all other eigenvalues are contained in $[-r, r]$. Then the decomposition in (A.5) can be written as

$$P = \sum_{i=1}^{m} \lambda_i \phi_i \psi_i^T + P_{\text{fast}} \tag{A.6}$$

The matrix P_{fast} contains the remaining eigenvalues and eigenvectors of P so that $P_{\text{fast}} \phi_i = 0$ for all $i = 1, \ldots, m$.

The evolution of a probability vector ν under the dynamics encoded by the transition matrix P can be written as

$$\nu^T P^n = \sum_{i=1}^{m} \lambda_i^n \langle \nu, \phi_i \rangle \psi_i^T + \nu^T P_{\text{fast}}^n. \tag{A.7}$$

In the long time limit (large n) the dominant contribution to the sum is the first term for the unit eigenvalue $\lambda_1 = 1$. Since all other eigenvalues have modulus smaller than one $\lim_{n \to \infty} \nu^T P^n = \pi$ for any probability vector ν, i.e. in the long time limit any initial ensemble will be distributed according to the equilibrium probabilities.

The slowly decaying terms correspond to slow processes representing dynamical rearrangements taking place in the ensemble while it relaxes to equilibrium [84]. To each dominant eigenvalue one can associate an relaxation time (implied timescale) for the transport of probability towards equilibrium

$$t_i = -\frac{\tau}{\log \lambda_i}. \tag{A.8}$$

The largest relaxation time t_2 governs the speed of convergence to equilibrium, i.e. in order to ensure that any equilibrium expectation value computed from a single molecular dynamics trajectory has converged the trajectory length has to be many times t_2 [84].

A.3 MEAN FIRST-PASSAGE TIMES BETWEEN META-STABLE REGIONS

The covered material can be found in many introductory books to stochastic processes, cf. [47].

For a stochastic process (X_t) on a state space Ω the first hitting time T_B of a set $B \subseteq \Omega$ is defined as

$$T_B = \inf\{t \geq 0 | X_t \in B\}. \tag{A.9}$$

The mean first passage time $\tau_{x,B}$ to the set B starting in state $x \in \Omega$ is the following expectation value

$$\tau_{x,B} = \mathbb{E}_x(T_B). \tag{A.10}$$

For a Markov chain on a finite state space $\Omega = \{1, \ldots, n\}$ with transition matrix $(p_{x,y})$ the mean first-passage time can be computed from the following system of equations,

$$\tau_{x,B} = \begin{cases} 0 & x \in B \\ 1 + \sum_{y \in \Omega} p_{x,y} \tau_{y,B} & x \notin B \end{cases} \tag{A.11}$$

Assuming that the chain has equilibrium distribution (π_x) we define the mean first-passage time $\tau_{A,B}$ from set A to set B as the μ-weighted average of all mean first-passage times to B when starting in a state $x \in A$,

$$\tau_{A,B} = \sum_{x \in A} \pi_x \tau_{x,B}. \tag{A.12}$$

Computing the mean first-passage time between two sets for a Markov chain on a finite state space with given transition matrix thus amounts to finding the equilibrium distribution together with the solution of a linear system of equations - both of which can be achieved using standard numerical linear algebra libraries.

A.4 COMMITTOR FUNCTIONS

Committor functions have been introduced in the context of Transition Path Theory [119] and are a central object for the characterization of transition processes between two meta-stable sets.

Let (X_t) again be a stochastic process on a state space Ω and let $A, B \subseteq \Omega$ be two meta-stable sets. The forward committor $q^{(+)}(x)$ is the probability that the process starting in x will reach the set B first, rather than the set A,

$$q^{(+)}(x) = \mathbb{P}_x(T_A < T_B). \tag{A.13}$$

Again T_S denotes the first hitting time of a set S.

For a Markov chain on a finite state space with transition matrix P the forward committor solves the following boundary value problem [64],

$$
\begin{array}{ll}
\sum_j l_{ij} q_j^{(+)} = 0 & i \in X \; (A \cup B) \\
q_i^{(+)} = 0 & i \in A \\
q_i^{(+)} = 1 & i \in B
\end{array}
\qquad . \qquad (A.14)
$$

$L = P - I$ is the corresponding generator matrix of the Markov chain.

Computing the committor for a finite state space again amounts to solving a linear system of equations.

TRANSITION MATRIX SAMPLING

We need the following elementary result.

Lemma 1 *Let X be a random variable with values in $\mathbb{R}^n$ having density $p_X(x)$ (w.r.t. the Lebesgue measure on $\mathbb{R}^n$). Let $Y = \Phi(X)$ with Φ an injective, continuous differentiable mapping with nonzero Jacobian $D\Phi$. Then the density for Y is*

$$p_Y(y) = p_X(\Phi^{-1}(y))|D\Phi^{-1}(y)|. \tag{B.1}$$

The factor $|D\Phi^{-1}(y)|$ is the determinant of the Jacobian of the inverse transformation.

B.1 REVERSIBLE SAMPLING

B.1.1 *Posterior*

The posterior (4.12) is a probability distribution on the lower triangular entries of the matrix X. In (4.13) we define a transformation between the entries of X and the entries of a new matrix V. The matrix X can be recovered from V using the inverse transformation,

$$x_{ij} = \frac{v_{ij}}{1 + v_{kl}}. \tag{B.2}$$

The Jacobian of Φ^{-1} is a block matrix

$$D\Phi^{-1}(v) = \begin{pmatrix} A(v) & B(v) \\ 0 & D(v) \end{pmatrix} \tag{B.3}$$

with diagonal matrix $A(v)$ with $n(n+1)/2 - 2$ diagonal entries $(1 + v_{kl})^{-1}$ and 1×1 block $D = (1 + v_{kl})^{-2}$. The determinant of a block upper triangular matrix can be conveniently computed as the product of the determinants of the diagonal blocks. The matrix $A(v)$ has only $n(n+1)/2 - 2$ diagonal entries since one value of x and v is completely determined by the normalization condition. The last row of the Jacobian corresponds to the variable $x_{kl} = v_{kl}/(1 + v_{kl})$.

The Jacobian determinant of (B.2) is

$$\left| D\Phi^{-1}(v) \right| = (1 + v_{kl})^{-n(n+1)/2}. \tag{B.4}$$

The posterior in new coordinates is then,

$$\mathbb{P}(V|C) \propto (1 + v_{kl})^{-(b_0 + \frac{n(n+1)}{2})} \prod_{i \geq j} v_{ij}^{b_{ij}} \prod_{i,j} \left(\frac{v_{ij}}{v_i} \right)^{c_{ij}}. \tag{B.5}$$

Expressions (4.16), (4.17) are the conditional distributions for V with density (B.5).

B.1.2   *Beta sampling for diagonal elements*

We show that sampling of the diagonal elements can be achieved by sampling from the Beta distribution. The transformation from s to v'_{kk} in (4.19) has Jacobian determinant,

$$|D\Phi^{-1}(s)| = \frac{v_{k,-k}}{(1-s)^2}. \tag{B.6}$$

For prior choice $b_{ij} = -1$ the transformed conditional, (4.16), is

$$\gamma(s|V) \propto (\frac{s}{1-s})^{c_{kk}-1}(1-s)^{-2}(1-s)^{c_k} = s^{c_{kk}-1}(1-s)^{c_k - c_{kk}-1} \tag{B.7}$$

This is a Beta distribution with parameters c_{kk} and $c_k - c_{kk}$.

B.1.3   *Gamma proposal for off-diagonal elements*

We show that matching the conditional for v_{kl}, (4.17), with a Gamma proposal results in parameters given in (4.21). We use the following representation for the conditional distribution

$$\gamma(v'_{kl}|V) \propto (v'_{kl})^{-1} \exp f(v'_{kl}) \tag{B.8a}$$
$$f(v'_{kl}) = (c_{kl} + c_{lk}) \log v'_{kl} - c_k \log(v_{k,-l} + v'_{kl}) - c_l \log(v_{l,-k} + v'_{kl}) \tag{B.8b}$$

and for the proposal density

$$q(v'_{kl}|V) \propto (v'_{kl})^{\alpha-1} e^{-\beta v'_{kl}} = (v'_{kl})^{-1} \exp \hat{f}(v'_{kl}) \tag{B.9a}$$
$$\hat{f}(v'_{kl}) = \alpha \log v'_{kl} - \beta v'_{kl}. \tag{B.9b}$$

To approximate $\gamma(v'_{kl}|V)$ by $q(v'_{kl}|V)$ we match first and second derivatives of f and $\hat{f}$ at the maximum point $\bar{v} = \arg\max f(v_{kl})$. The necessary condition for a maximum point, $f'(\bar{v}) = 0$, results in a quadratic equation with solution

$$\bar{v} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}. \tag{B.10}$$

The second solution, (B.10) with negative sign in front of the square root, can be safely excluded since $\bar{v}$ is required to be non-negative. The parameters a, b, c are given by

$$a = c_k + c_l - c_{kl} - c_{lk} \tag{B.11a}$$
$$b = c_k v_{l,-k} + c_l v_{k,-l} - (c_{kl} + c_{lk})(v_{k,-l} + v_{l,-k}) \tag{B.11b}$$
$$c = -(c_{kl} + c_{lk})v_{k,-l}v_{l,-k}. \tag{B.11c}$$

Matching first and second derivative of f and $\hat{f}$ at $\bar{v}$ leads to the following linear system for the parameters α and β,

$$\frac{\alpha}{\bar{v}} - \beta = 0 \tag{B.12a}$$
$$-\frac{\alpha}{\bar{v}^2} = f''(\bar{v}). \tag{B.12b}$$

The solution to (B.12) is

$$\alpha = -f''(\bar{v})\bar{v}^2 \tag{B.13a}$$

$$\beta = -f''(\bar{v})\bar{v}. \tag{B.13b}$$

The parameter h in (4.21) is the value of the second derivative at the maximum,

$$h = f''(\bar{v}) = -\frac{c_{kl} + c_{lk}}{\bar{v}^2} + \frac{c_k}{(v_{k,-l} + \bar{v})^2} + \frac{c_l}{(v_{k,-l} + \bar{v})^2}.$$

The log acceptance probability in (4.22) can be obtained if one inserts (B.8a) and (B.9a) into (4.18),

$$\log p_{acc} = f(v'_{kl}) - f(v_{kl}) + \hat{f}(v_{kl}) - \hat{f}(v'_{kl}). \tag{B.14}$$

Plugging in (B.8b) and (B.9b) we obtain (4.22).

B.1.4 *Logspace random walk*

We show that the acceptance probability for the logspace random walk is given by (4.23).

The proposal distribution for the Gaussian random walk on $\log v_{kl}$ is

$$q(v'_{kl}|V) \propto (v'_{kl})^{-1} \exp \frac{(\log v'_{kl} - \log v_{kl})^2}{2} = (v'_{kl})^{-1} \exp \hat{f}(v'_{kl}) \tag{B.15a}$$

$$\hat{f}(v'_{kl}) = \frac{(\log v'_{kl} - \log v_{kl})^2}{2}. \tag{B.15b}$$

The function $\hat{f}$ is invariant with respect to an exchange of v'_{kl} and v_{kl} so that the acceptance probability depends only on the value of the conditional,

$$\log p_{acc} = f(v'_{kl}) - f(v_{kl}). \tag{B.16}$$

B.2 REVERSIBLE SAMPLING WITH FIXED STATIONARY VECTOR

B.2.1 *Conditional expectation and likelihood*

We show that for $c_{kk} = 0$ the choice $b_{kk} = 0$ for the diagonal prior parameter matches the expectation of the conditional for x_{kl}, (4.27), and the maximum point of the conditional likelihood function.

For $c_{kk} = 0$ and $b_{kk} = 0$ the conditional is

$$\gamma(x'_{kl}|X) \propto x_{kl}^{c_{kl}+c_{lk}+b_{kl}}(\Delta_{lk} - x_{kl})^{c_{ll}+b_{ll}} \tag{B.17}$$

The density (B.17) is a Beta distribution with parameters $\alpha = c_{kl} + c_{lk} + b_{kl} + 1$ and $\beta = c_{ll} + b_{ll} + 1$ for the scaled variable $y_{kl} = x_{kl}/\Delta_{lk}$. The expectation value for the Beta distribution is $\alpha/(\alpha + \beta)$ so that

$$\mathbb{E}(x_{kl}) = \Delta_{lk} \frac{c_{kl} + c_{lk} + b_{kl} + 1}{c_{kl} + c_{lk} + b_{kl} + 1 + c_{ll} + b_{ll} + 1}. \tag{B.18}$$

The likelihood function for x_{kl}, (B.17) with $b_{kl} = 0$ and $b_{ll} = 0$, is also a Beta distribution. The maximum of the Beta distribution is $(\alpha - 1)/(\alpha + \beta - 2)$ so that

$$\hat{x}_{kl} = \frac{c_{kl} + c_{lk}}{c_{kl} + c_{lk} + c_{ll}} \Delta_{kl} \tag{B.19}$$

The expressions on the right hand side of (B.18) and (B.19) are equal for $b_{kl} = b_{ll} = -1$.

B.2.2   *Conditional distribution*

We show that the conditional distribution for the element v_{kl} is given by (4.30) with parameters (4.31).

The transformation from x_{kl} to v in (4.29) has inverse

$$x_{kl} = \frac{v}{1 + v} \Delta_{kl}. \tag{B.20}$$

The Jacobian determinant of the inverse is

$$|D\Phi^{-1}(v)| = \frac{\Delta_{kl}}{(1 + v)^2}. \tag{B.21}$$

The conditional (4.27) in new variables is then

$$\gamma(v|X) \propto v^{c_{kl} + c_{lk} + b_{kl}} (1 + v)^{-c_{kl} - c_{lk} - b_{kl} - c_{kk} - b_{kk} - c_{ll} - b_{ll} - 2} \\ (\Delta_{lk}(1 + v) - \Delta_{kl} v)^{c_{ll} + b_{ll}}. \tag{B.22}$$

Defining $s = \Delta_{kl}/\Delta_{lk}$ we obtain (4.30) with parameters in (4.31).

B.2.3   *Gamma proposal*

We show that matching the conditional for v, (4.30), results in parameters given in (4.32).

Again, we represent the conditional (4.30) as in (B.8a). The function f is now

$$f(v) = (a_1 - 1) \log v + a_2 \log(r + v) - (a_1 + a_2 + a_2 + 2) \log(1 + v) \tag{B.23}$$

The matching procedure is identical to the one outlined in Section B.1.3 and results in parameters given in (4.32).

The acceptance probability in (4.33) can be computed from (B.14) using f in (B.23) and $\hat{f}$ from (B.9b).
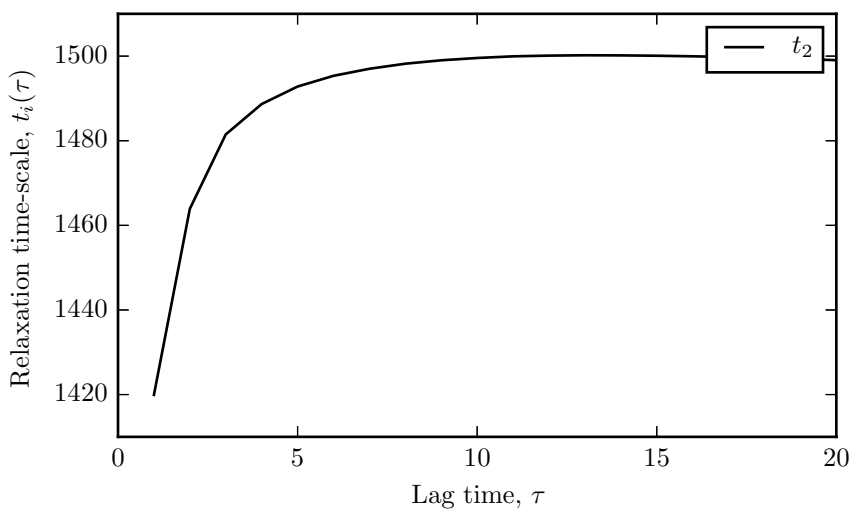
B.2.4 *Logspace random walk*

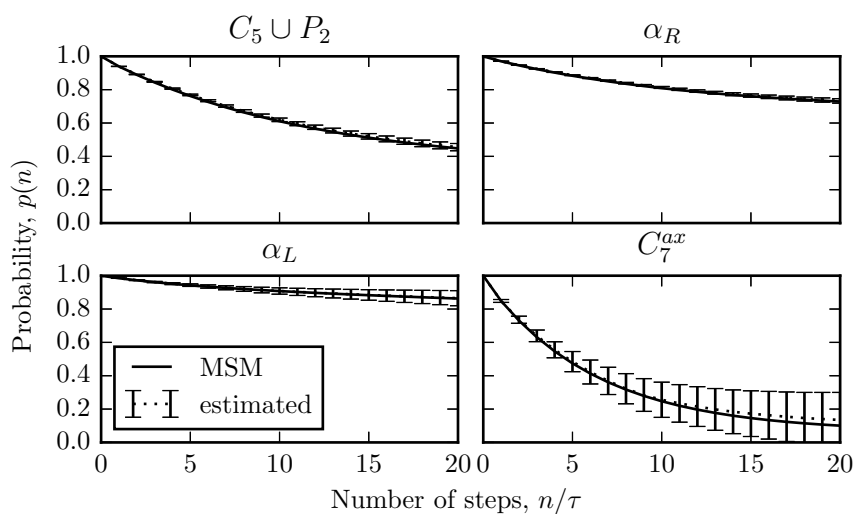We show that the acceptance probability for the logspace random walk is given by (4.34).

The proposal distribution is identical to (B.15). The acceptance probability in (4.34) can be computed from (B.16) with f in (B.23).

Figure C.1: MSM validation for alanine dipeptide ϕ, ψ dihedral angle data. a) Implied timescale test. Convergence of the largest relaxation time-scale, t_2, indicates a good Markov model fit, i.e. the slow eigenfunction of the associated dynamical operator are well approximated. b) The Chapman-Kolmogorov test validates the Markov assumption by comparing the evolution of self-transition probabilities predicted by the MSM parameterized at lagtime τ with direct estimates from the data at larger lagtimes $n\tau$.

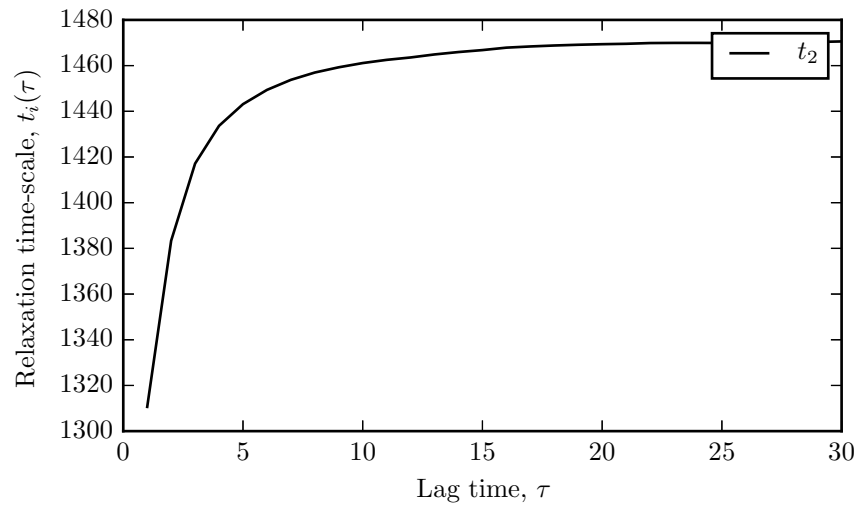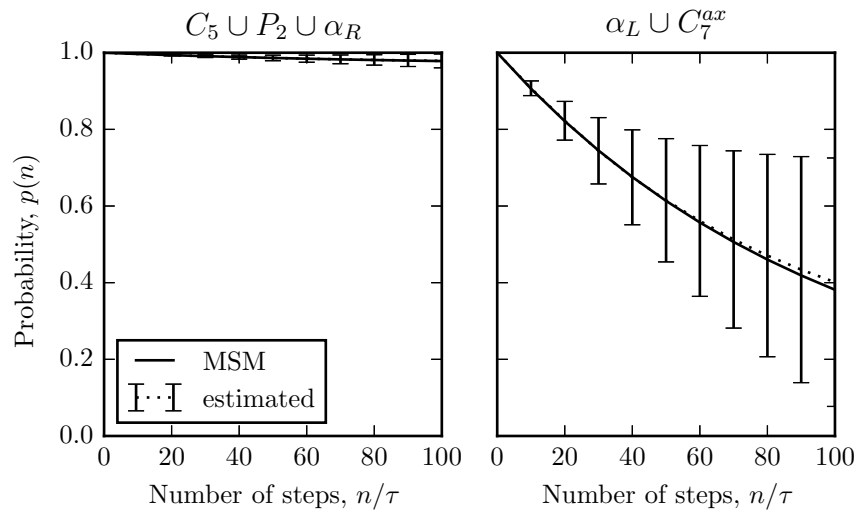(a)

(b)

Figure C.2: MSM validation for alanine dipeptide ϕ dihedral angle data. a) Implied timescale test. Convergence of the largest relaxation time-scale, t_2, indicates a good Markov model fit, i.e. the slow eigenfunction of the associated dynamical operator are well approximated. b) The Chapman-Kolmogorov test validates the Markov assumption by comparing the evolution of self-transition probabilities predicted by the MSM parameterized at lagtime τ with direct estimates from the data at larger lagtimes $n\tau$.

(a)

(b)

Figure C.3: MSM validation for vesicle model. a) Implied timescale test. Convergence of the largest relaxation time-scale, t_2, indicates a good Markov model fit, i.e. the slow eigenfunction of the associated dynamical operator are well approximated. b) The Chapman-Kolmogorov test validates the Markov assumption by comparing the evolution of self-transition probabilities predicted by the MSM parameterized at lagtime τ wit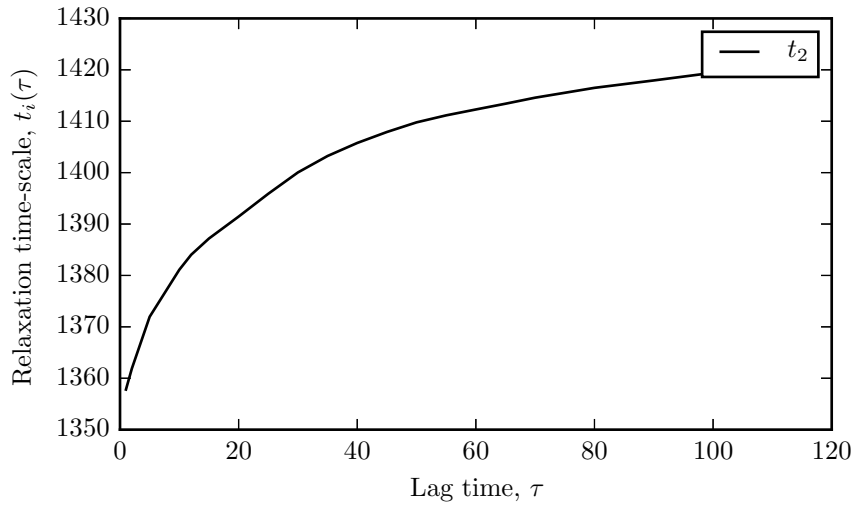h d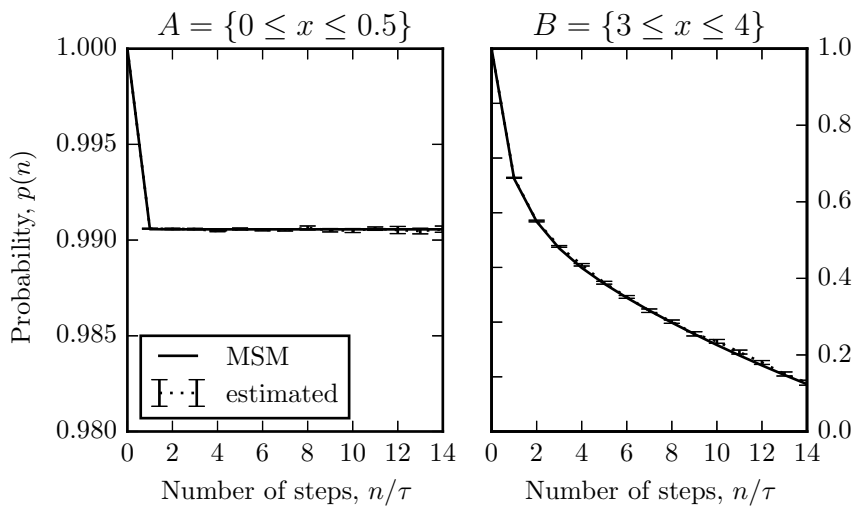irect estimates from the data at larger lagtimes $n\tau$. Values were obtained from an ensemble of short trajectories starting in the high energy region utilizing the equilibrium distribution in the estimation of the MLE transition matrix, cf. (3.26).

[1] A. G. Anderson and J. Hermans. "Microfolding: conformational probability map for the alanine dipeptide in water from molecular dynamics simulations." In: *Proteins* 3.4 (1988), pp. 262–265. URL: http://dx.doi.org/10.1002/prot.340030408.

[2] T. W. Anderson and L. A. Goodman. "Statistical Inference about Markov Chains." In: *Ann. Math. Statist.* 28.1 (1957), pp. 89–110. URL: http://dx.doi.org/10.1214/aoms/1177707039.

[3] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. "Dynamics of ligand binding to myoglobin." In: *Biochemistry* 14.24 (1975), pp. 5355–5373. URL: http://dx.doi.org/10.1021/bi00695a021.

[4] S. Bacallado, J. D. Chodera, and V. Pande. "Bayesian comparison of Markov models of molecular dynamics with detailed balance constraint." In: *J. Chem. Phys.* 131.4, 045106 (2009). URL: http://scitation.aip.org/content/aip/journal/jcp/131/4/10.1063/1.3192309.

[5] C. H. Bennett. "Efficient estimation of free energy differences from Monte Carlo data." In: *J. Comput. Phys.* 22.2 (1976), pp. 245 –268. URL: http://www.sciencedirect.com/science/article/pii/0021999176900784.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. "The Protein Data Bank." In: *Nucleic Acids Res.* 28.1 (2000), pp. 235–242. URL: http://nar.oxfordjournals.org/content/28/1/235.abstract.

[7] R. B. Best and G. Hummer. "Coordinate-dependent diffusion in protein folding." In: *Proc. Natl. Acad. Sci. USA* 107.3 (2010), pp. 1088–1093. URL: http://www.pnas.org/content/107/3/1088.abstract.

[8] A. Bittracher, P. Koltai, and O. Junge. "Pseudogenerators of Spatial Transfer Operators." In: *SIAM J. Appl. Dyn. Syst.* 14.3 (2015), pp. 1478–1517. URL: http://dx.doi.org/10.1137/14099872X.

[9] A. Bittracher, C. Hartmann, O. Junge, and P. Koltai. *Pseudo generators for under-resolved molecular dynamics.* 2015. URL: http://arxiv.org/abs/1502.01191.

[10]   P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark." In: *Annu. Rev. Phys. Chem.* 53.1 (2002), pp. 291–318. URL: http://dx.doi.org/10.1146/annurev.physchem.53.082301.113146.

[11]   G. R. Bowman, V. S. Pande, and F. Noé. *An introduction to markov state models and their application to long timescale molecular simulation.* Vol. 797. Springer Science & Business Media, 2013.

[12]   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." In: *J. Chem. Phys.* 131.12, 124101 (2009), pp. –.

[13]   S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[14]   I. Buch, T. Giorgino, and G. de Fabritiis. "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations." In: *Proc. Natl. Acad. Sci. USA* 108.25 (2011), pp. 10184–10189. URL: http://www.pnas.org/content/108/25/10184.abstract.

[15]   J. R. Bunch and L. Kaufman. "Some stable methods for calculating inertia and solving symmetric linear systems." In: *Math. Comput.* (1977), pp. 163–179. URL: http://www.ams.org/journals/mcom/1977-31-137/S0025-5718-1977-0428694-0.

[16]   T. F. C. Chan and T. P. Mathew. "The Interface Probing Technique in Domain Decomposition." In: *SIAM J. Matrix Anal. A.* 13.1 (1992), pp. 212–238. URL: http://dx.doi.org/10.1137/0613018.

[17]   S. Chib and E. Greenberg. *Understanding the Metropolis-Hastings Algorithm.* 1995. URL: http://www2.stat.duke.edu/~scs/Courses/Stat376/Papers/Basic/ChibGreenberg1995.pdf.

[18]   J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill. "Long-time protein folding dynamics from short-time molecular dynamics simulations." In: *Multiscale Model. Simul.* 5.4 (2006), pp. 1214–1226. URL: http://dx.doi.org/10.1137/06065146X.

[19]   J. B. Clarage, T Romo, B. K. Andrews, B. M. Pettitt, and G. N. Phillips. "A sampling problem in molecular dynamics simulations of macromolecules." In: *Proc. Natl. Acad. Sci. USA* 92.8 (1995), pp. 3288–3292. URL: http://www.pnas.org/content/92/8/3288.abstract.

[20]   X. Daura, A. E. Mark, and W. F. van Gunsteren. "Peptide folding simulations: no solvent required?" In: *Comput. Phys. Commun.* 123.1–3 (1999), pp. 97 –102. URL: http://www.sciencedirect.com/science/article/pii/S0010465599002611.

[21]   C. Dellago, P. G. Bolhuis, and D. Chandler. "Efficient transition path sampling: Application to Lennard-Jones cluster rearrangements." In: *J. Chem. Phys.* 108.22 (1998), pp. 9236–9245. URL: http://scitation.aip.org/content/aip/journal/jcp/108/22/10.1063/1.476378.

[22]   J. L. Denny and A. L. Wright. "On tests for Markov dependence." In: *Z. Wahrscheinlichkeit.* 43.4 (1978), pp. 331–338. URL: http://dx.doi.org/10.1007/BF00534766.

[23]   P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. "Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains." In: *Linear Algebra Appl.* 315.1–3 (2000), pp. 39 –59. URL: http://www.sciencedirect.com/science/article/pii/S0024379500000951.

[24]   P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. "Computational Molecular Dynamics: Challenges, Methods, Ideas: Proceedings of the 2nd International Symposium on Algorithms for Macromolecular Modelling, Berlin, May 21–24, 1997." In: ed. by P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, and R. D. Skeel. Berlin, Heidelberg: Springer, 1999. Chap. Computation of Essential Molecular Dynamics by Subdivision Techniques, pp. 98–115. URL: http://dx.doi.org/10.1007/978-3-642-58360-5_5.

[25]   L. Devroye. *Non-Uniform Random Variate Generation.* Springer, 1986. URL: http://luc.devroye.org/rnbookindex.html.

[26]   W.-N. Du, K. A. Marino, and P. G. Bolhuis. "Multiple state transition interface sampling of alanine dipeptide in explicit solvent." In: *J. Chem. Phys.* 135.14 (2011), p. 145102. URL: http://scitation.aip.org/content/aip/journal/jcp/135/14/10.1063/1.3644344.

[27]   P. Eastman et al. "OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation." In: *J. Chem. Theory Comput.* 9.1 (2013), pp. 461–469. URL: http://dx.doi.org/10.1021/ct300857j.

[28]   B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[29]   R Elber and M Karplus. "Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin." In: *Science* 235.4786 (1987), pp. 318–321. URL: http://science.sciencemag.org/content/235/4786/318.

[30]   T. S. van Erp, D. Moroni, and P. G. Bolhuis. "A novel path sampling method for the calculation of rate constants." In: *J. Chem. Phys.* 118.17 (2003). URL: http://scitation.aip.org/content/aip/journal/jcp/118/17/10.1063/1.1562614.

[31]   H. Eyring. "The Activated Complex in Chemical Reactions." In: *J. Chem. Phys.* 3.2 (1935), pp. 107–115. URL: http://scitation.aip.org/content/aip/journal/jcp/3/2/10.1063/1.1749604.

[32]   F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems.* Springer Science & Business Media, 2007.

[33]   A. K. Faradjian and R. Elber. "Computing time scales from reaction coordinates by milestoning." In: *J. Chem. Phys.* 120.23 (2004), pp. 10880–10889. URL: http://scitation.aip.org/content/aip/journal/jcp/120/23/10.1063/1.1738640.

[34]   T. S. Ferguson. *A Course in Large Sample Theory.* Chapmann & Hall, 1996.

[35]   A. M. Ferrenberg and R. H. Swendsen. "Optimized Monte Carlo data analysis." In: *Phys. Rev. Lett.* 63 (1989), pp. 1195–1198. URL: http://link.aps.org/doi/10.1103/PhysRevLett.63.1195.

[36]   H. Frauenfelder and G. Petsko. "Structural dynamics of liganded myoglobin." In: *Biophys. J.* 32.1 (1980), pp. 465 –483. ISSN: 0006-3495. URL: http://www.sciencedirect.com/science/article/pii/S0006349580849848.

[37]   C. J. Geyer. "Practical Markov Chain Monte Carlo." In: *Statist. Sci.* 7.4 (1992), pp. 473–483. URL: http://dx.doi.org/10.1214/ss/1177011137.

[38]   D. S. Goodsell. *HIV Reverse Transcriptase.* 2002. URL: http://dx.doi.org/10.2210/rcsb_pdb/mom_2002_9.

[39]   D. S. Goodsell. *Major Histocompatibility Complex.* 2005. URL: http://dx.doi.org/10.2210/rcsb_pdb/mom_2005_2.

[40]   D. S. Goodsell. *Adrenergic Receptors.* 2008. URL: http://dx.doi.org/10.2210/rcsb_pdb/mom_2008_4.

[41]   D. S. Goodsell, S. Dutta, C. Zardecki, M. Voigt, H. M. Berman, and S. K. Burley. "The RCSB PDB "Molecule of the Month": Inspiring a Molecular View of Biology." In: *PLoS Biol.* 13.5 (2015), pp. 1–12. URL: http://dx.doi.org/10.1371/journal.pbio.1002140.

[42]   H. Grubmüller. "Predicting slow structural transitions in macromolecular systems: Conformational flooding." In: *Phys. Rev. E* 52 (1995), pp. 2893–2906. URL: http://link.aps.org/doi/10.1103/PhysRevE.52.2893.

[43] C. Hartmann, J. C. Latorre, and G. Ciccotti. "On two possible definitions of the free energy for collective variables." In: *Eur. Phys. J. Special Topics* 200.1 (2011), pp. 73–89. URL: http://dx.doi.org/10.1140/epjst/e2011-01519-7.

[44] C. Hartmann and C. Schütte. "Efficient rare event simulation by optimal nonequilibrium forcing." In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.11 (2012), P11004. URL: http://stacks.iop.org/1742-5468/2012/i=11/a=P11004.

[45] H Hartmann, F Parak, W Steigemann, G. A. Petsko, D. R. Ponzi, and H Frauenfelder. "Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K." In: *Proc. Natl. Acad. Sci. USA* 79.16 (1982), pp. 4967–4971. URL: http://www.pnas.org/content/79/16/4967.abstract.

[46] N. S. Hinrichs and V. S. Pande. "Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics." In: *J. Chem. Phys.* 126 (2007), p. 244101. URL: http://scitation.aip.org/content/aip/journal/jcp/126/24/10.1063/1.2740261.

[47] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to stochastic processes.* Waveland Press, 1986.

[48] W. Huisinga and B. Schmidt. "Metastability and dominant eigenvalues of transfer operators." In: *New Algorithms for Macromolecular Simulation.* Springer, 2006, pp. 167–182.

[49] S. Ji, Y. Xue, and L. Carin. "Bayesian Compressive Sensing." In: *Signal Processing, IEEE Transactions on* 56.6 (2008), pp. 2346–2356. URL: http://dx.doi.org/10.1109/TSP.2007.914345.

[50] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. "Comparison of simple potential functions for simulating liquid water." In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935. URL: http://scitation.aip.org/content/aip/journal/jcp/79/2/10.1063/1.445869.

[51] J. Kang, Y. Cao, D. P. Word, and C. Laird. "An interior-point method for efficient solution of block-structured {NLP} problems using an implicit Schur-complement decomposition." In: *Comput. Chem. Eng.* 71 (2014), pp. 563 –573. URL: http://www.sciencedirect.com/science/article/pii/S0098135414002798.

[52] L. E. Kay, D. A. Torchia, and A. Bax. "Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease." In: *Biochemistry* 28.23 (1989), pp. 8972–8979. URL: http://dx.doi.org/10.1021/bi00449a003.

[53] H. Keller and P. G. Debrunner. "Evidence for Conformational and Diffusional Mean Square Displacements in Frozen Aqueous Solution of Oxymyoglobin." In: *Phys. Rev. Lett.* 45 (1 1980), pp. 68–71. URL: http://link.aps.org/doi/10.1103/PhysRevLett.45.68.

[54] C. Kipnis and S. R. S. Varadhan. "Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions." In: *Comm. Math. Phys.* 104.1 (1986), pp. 1–19. URL: http://projecteuclid.org/euclid.cmp/1104114929.

[55] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Vol. 23. Springer, 1992.

[56] H. Kramers. "Brownian motion in a field of force and the diffusion model of chemical reactions." In: *Physica* 7.4 (1940), pp. 284–304. URL: http://www.sciencedirect.com/science/article/pii/S0031891440900982.

[57] S. Kumar, J. Rosenberg, D. Bouzida, R. Swendsen, and P. Kollman. "Multidimensional free-energy calculations using the weighted histogram analysis method." In: *J. Comput. Chem.* 16.11 (1995), pp. 1339–1350. URL: http://dx.doi.org/10.1002/jcc.540161104.

[58] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method." In: *J. Comput. Chem.* 13.8 (1992), pp. 1011–1021. URL: http://dx.doi.org/10.1002/jcc.540130812.

[59] A. Laio and M. Parrinello. "Escaping free-energy minima." In: *Proc. Natl. Acad. Sci. USA* 99.20 (2002), pp. 12562–12566. URL: http://www.pnas.org/content/99/20/12562.abstract.

[60] B. Leimkuhler and C. Matthews. *Molecular Dynamics: with deterministic and stochastic numerical methods.* Vol. 39. Springer, 2015.

[61] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, 2009.

[62] 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." In: *Proteins* 78.8 (2010), pp. 1950–1958. URL: http://dx.doi.org/10.1002/prot.22711.

[63] P. Metzner, F. Noé, and C. Schütte. "Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data." In: *Phys. Rev. E* 80 (2009), p. 021106. URL: http://link.aps.org/doi/10.1103/PhysRevE.80.021106.

[64]   P. Metzner, C. Schütte, and E. Vanden-Eijnden. "Transition Path Theory for Markov Jump Processes." In: *Multiscale Model. Sim.* 7.3 (2009), pp. 1192–1219. URL: http://dx.doi.org/10.1137/070699500.

[65]   P. Metzner, M. Weber, and C. Schütte. "Observation uncertainty in reversible Markov chains." In: *Phys. Rev. E* 82 (3 2010), p. 031114. URL: http://link.aps.org/doi/10.1103/PhysRevE.82.031114.

[66]   L. Molgedey and H. G. Schuster. "Separation of a mixture of independent signals using time delayed correlations." In: *Phys. Rev. Lett.* 72 (23 1994), pp. 3634–3637. URL: http://link.aps.org/doi/10.1103/PhysRevLett.72.3634.

[67]   B Montgomery Pettitt and M. Karplus. "The potential of mean force surface for the alanine dipeptide in aqueous solution: a theoretical approach." In: *Chem. Phys. Lett.* 121.3 (1985), pp. 194–201. URL: http://www.sciencedirect.com/science/article/pii/0009261485855093.

[68]   L. Monticelli and D. P. Tieleman. "Force fields for classical molecular dynamics." In: *Methods Mol. Biol.* 924 (2013), pp. 197–213. URL: http://dx.doi.org/10.1007/978-1-62703-017-5_8.

[69]   A. Nielsen. "Von Femtosekunden zu Minuten." MA thesis. FU Berlin, 2012. URL: http://www.zib.de/nielsen/Masterarbeit.pdf.

[70]   F. Noé, I. Horenko, C. Schütte, and J. C. Smith. "Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States." In: *J. Chem. Phys.* 126 (2007), p. 155102. URL: http://scitation.aip.org/content/aip/journal/jcp/126/15/10.1063/1.2714539.

[71]   F. Noé. *Statistical inefficiency of Markov model count matrices.* URL: http://publications.mi.fu-berlin.de/1699/1/autocorrelation_counts.pdf.

[72]   F. Noé. "Transition Networks: Computational Methods for the Comprehensive Analysis of Complex Rearrangements in Proteins." PhD thesis. University of Heidelberg, 2006. URL: http://publications.mi.fu-berlin.de/433/.

[73]   F. Noé. "Probability distributions of molecular observables computed from Markov models." In: *J. Chem. Phys.* 128.24, 244103 (2008). URL: http://scitation.aip.org/content/aip/journal/jcp/128/24/10.1063/1.2916718.

[74]   F. Noé and F. Nüske. "A Variational Approach to Modeling Slow Processes in Stochastic Dynamical Systems." In: *Multiscale Model. Sim.* 11.2 (2013), pp. 635–655. URL: http://dx.doi.org/10.1137/110858616.

[75]   F. Noé, F. Ille, J. C. Smith, and S. Fischer. "Automated computation of low-energy pathways for complex rearrangements in proteins: Application to the conformational switch of Ras p21." In: *Proteins: Structure, Function, and Bioinformatics* 59.3 (2005), pp. 534–544. URL: http://dx.doi.org/10.1002/prot.20422.

[76]   F. Noé, M. Oswald, G. Reinelt, S. Fischer, and J. C. Smith. "Computing Best Transition Pathways in High-Dimensional Dynamical Systems: Application to the AlphaL-Beta-AlphaR Transitions in Octaalanine." In: *Multiscale Model. Sim.* 5.2 (2006), pp. 393–419. URL: http://dx.doi.org/10.1137/050641922.

[77]   F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations." In: *Proc. Natl. Acad. Sci. USA* 106.45 (2009), pp. 19011–19016. URL: http://www.pnas.org/content/106/45/19011.abstract.

[78]   J. R. Norris. *Markov Chains.* Cambridge University Press, 1998.

[79]   F. Nüske, B. G. Keller, G. Pérez-Hern'andez, A. S. J. S. Mey, and F. Noé. "Variational Approach to Molecular Kinetics." In: *J. Chem. Theory Comput.* 10.4 (2014), pp. 1739–1752. URL: http://dx.doi.org/10.1021/ct4009156.

[80]   F. Nüske, R. Schneider, F. Vitalini, and F. Noé. "Variational tensor approach for approximating the rare-event kinetics of macromolecular systems." In: *J. Chem. Phys.* 144.5, 054105 (2016). URL: http://scitation.aip.org/content/aip/journal/jcp/144/5/10.1063/1.4940774.

[81]   C. C. Paige and M. A. Saunders. "Solution of sparse indefinite systems of linear equations." In: *SIAM J. Numer. Anal.* 12.4 (1975), pp. 617–629. URL: http://dx.doi.org/10.1137/0712047.

[82]   A. Pan and B. Roux. "Building Markov state models along pathways to determine free energies and rates of transitions." In: *J. Chem. Phys.* 129.6 (2008). URL: http://scitation.aip.org/content/aip/journal/jcp/129/6/10.1063/1.2959573.

[83]   G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé. "Identification of slow molecular order parameters for Markov model construction." In: *J. Chem. Phys.* 139.1, 015102 (2013). URL: http://scitation.aip.org/content/aip/journal/jcp/139/1/10.1063/1.4811489.

[84]   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." In: *J. Chem. Phys.* 134.17, 174105 (2011). URL: http://scitation.aip.org/content/aip/journal/jcp/134/17/10.1063/1.3565032.

[85] W. W. Priscilla E. Greenwood. "Reversible Markov Chains and Optimality of Symmetrized Empirical Estimators." In: *Bernoulli* 5.1 (1999), pp. 109–123. URL: http://www.jstor.org/stable/3318615.

[86] D. Ralph and S. J. Wright. "Superlinear Convergence of an Interior-Point Method Despite Dependent Constraints." In: *Math. Oper. Res.* 25.2 (2000), pp. 179–194. URL: http://www.jstor.org/stable/3690576.

[87] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2004.

[88] S. Röblitz. "Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics." PhD thesis. FU Berlin, 2008. URL: https://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/1131.

[89] S. Röblitz and M. Weber. "Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification." In: *Advances in Data Analysis and Classification* 7.2 (2013), pp. 147–179. URL: http://dx.doi.org/10.1007/s11634-013-0134-6.

[90] E. Rosta and G. Hummer. "Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model." In: *J. Chem. Theory Comput.* 11.1 (2015), pp. 276–285. URL: http://dx.doi.org/10.1021/ct500719p.

[91] M. Sarich, F. Noé, and C. Schütte. "On the Approximation Quality of Markov State Models." In: *Multiscale Model. Sim.* 8.4 (2010), pp. 1154–1177. URL: http://dx.doi.org/10.1137/090764049.

[92] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, 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." In: *J. Chem. Theory Comput.* (2015). URL: http://dx.doi.org/10.1021/acs.jctc.5b00743.

[93] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan. "Extracting Markov Models of Peptide Conformational Dynamics from Simulation Data." In: *J. Chem. Theory Comp.* 1.4 (2005), pp. 515–526. URL: http://dx.doi.org/10.1021/ct050020x.

[94] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. "A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo." In: *J. Comput. Phys.* 151.1 (1999), pp. 146 –168. URL: http://www.sciencedirect.com/science/article/pii/S0021999199962319.

[95]   C. Schütte, W. Huisinga, and P. Deuflhard. *Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems*. Tech. rep. Konrad-Zuse-Zentrum Berlin, 1999. URL: http://publications.mi.fu-berlin.de/129/1/SC-99-36.pdf.

[96]   C. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. AMS, 2014.

[97]   C. R. Schwantes and V. S. Pande. "Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9." In: *J. Chem. Theory Comput.* 9.4 (2013), pp. 2000–2009. URL: http://dx.doi.org/10.1021/ct300878a.

[98]   M. Shirts and V. S. Pande. "COMPUTING: Screen Savers of the World Unite!" In: *Science* 290.5498 (2000), pp. 1903–1904. URL: http://science.sciencemag.org/content/290/5498/1903.

[99]   M. R. Shirts and J. D. Chodera. "Statistically optimal analysis of samples from multiple equilibrium states." In: *J. Chem. Phys.* 129.12, 124105 (2008). URL: http://scitation.aip.org/content/aip/journal/jcp/129/12/10.1063/1.2978177.

[100]   M. R. Shirts and V. S. Pande. "Mathematical Analysis of Coupled Parallel Simulations." In: *Phys. Rev. Lett.* 86 (22 2001), pp. 4983–4987. URL: http://link.aps.org/doi/10.1103/PhysRevLett.86.4983.

[101]   N. Singhal and V. S. Pande. "Error analysis and efficient sampling in Markovian state models for molecular dynamics." In: *J. Chem. Phys.* 123.20, 204909 (2005). URL: http://scitation.aip.org/content/aip/journal/jcp/123/20/10.1063/1.2116947.

[102]   N. Singhal, C. D. Snow, and V. S. Pande. "Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin." In: *J. Chem. Phys.* 121.1 (2004), pp. 415–425. URL: http://scitation.aip.org/content/aip/journal/jcp/121/1/10.1063/1.1738647.

[103]   E. J. Sorin and V. S. Pande. "Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations." In: *Biophys. J.* 88.4 (2005), pp. 2472 –2493. ISSN: 0006-3495. URL: http://dx.doi.org/10.1529/biophysj.104.051938.

[104]   Y. Sugita and Y. Okamoto. "Replica-exchange molecular dynamics method for protein folding." In: *Chem. Phys. Lett.* 314.1 (1999), pp. 141–151. URL: http://www.sciencedirect.com/science/article/pii/S0009261499011239.

[105] W. C. Swope, J. W. Pitera, and F. Suits. "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory†." In: *J. Phys. Chem. B* 108.21 (2004), pp. 6571–6581. URL: http://dx.doi.org/10.1021/jp037421y.

[106] Z. Tan. "On a Likelihood Approach for Monte Carlo Integration." In: *J. Amer. Statist. Assoc.* 99.468 (2004), pp. 1027–1036. URL: http://www.jstor.org/stable/27590482.

[107] Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy. "Theory of binless multi-state free energy estimation with applications to protein-ligand binding." In: *J. Chem. Phys.* 136.14, 144102 (2012). URL: http://scitation.aip.org/content/aip/journal/jcp/136/14/10.1063/1.3701175.

[108] P. Tiwary and M. Parrinello. "From Metadynamics to Dynamics." In: *Phys. Rev. Lett.* 111 (23 2013), p. 230602. URL: http://link.aps.org/doi/10.1103/PhysRevLett.111.230602.

[109] D. J. Tobias and C. L. Brooks III. "Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results." In: *J. Phys. Chem.* 96.9 (1992), pp. 3864–3870. URL: http://dx.doi.org/10.1021/acs.jctc.5b00952.

[110] G. Torrie and J. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." In: *J. Comp. Phys.* 23.2 (1977), pp. 187–199. URL: http://www.sciencedirect.com/science/article/pii/0021999177901218.

[111] S. Trebst and M. Troyer. "Ensemble optimization techniques for classical and quantum systems." In: *Computer Simulations in Condensed Matter: From Materials to Chemical Biology. Volume 1*. Ed. by M. Ferrario, G. Ciccotti, and K. Binder. Springer, 2006.

[112] B. Trendelkamp-Schroer, M. Scherer, F. Paul, and F. Noé. *msmtools - A Python package for estimation and analysis of discrete space Markov chains via Markov state models (MSM)*. URL: https://github.com/markovmodel/msmtools.

[113] B. Trendelkamp-Schroer and F. Noé. "Efficient Estimation of Rare-Event Kinetics." In: *Phys. Rev. X* 6 (1 2016), p. 011009. URL: http://link.aps.org/doi/10.1103/PhysRevX.6.011009.

[114] B. Trendelkamp-Schroer, H. Wu, and F. Noé. *Reversible Markov chain estimation using convex-concave programming*. 2016. URL: http://arxiv.org/abs/1603.01640.

[115] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. "Estimation and uncertainty of reversible Markov models." In: *J. Chem. Phys.* 143.17 (2015), p. 174101. URL: http://scitation.aip.org/content/aip/journal/jcp/143/17/10.1063/1.4934536.

[116] M. Tuckerman, B. J. Berne, and G. J. Martyna. "Reversible multiple time scale molecular dynamics." In: *J. Chem. Phys.* 97.3 (1992), pp. 1990–2001. URL: http://scitation.aip.org/content/aip/journal/jcp/97/3/10.1063/1.463137.

[117] P. J. Tummino and R. A. Copeland. "Residence time of receptor-ligand complexes and its effect on biological function." In: *Biochemistry* 47.20 (2008), pp. 5481–5492. URL: http://pubs.acs.org/doi/abs/10.1021/bi8002023.

[118] W. F. Van Gunsteren and H. J. C. Berendsen. "A Leap-frog Algorithm for Stochastic Dynamics." In: *Molecular Simulation* 1.3 (1988), pp. 173–185. URL: http://dx.doi.org/10.1080/08927028808080941.

[119] E. Vanden-Eijnden. "Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1." In: ed. by M. Ferrario, G. Ciccotti, and K. Binder. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. Chap. Transition Path Theory, pp. 453–493. URL: http://dx.doi.org/10.1007/3-540-35273-2_13.

[120] E. Vanden-Eijnden and J. Weare. "Rare Event Simulation of Small Noise Diffusions." In: *Communications on Pure and Applied Mathematics* 65.12 (2012), pp. 1770–1803. URL: http://dx.doi.org/10.1002/cpa.21428.

[121] Y. Vardi. "Empirical Distributions in Selection Bias Models." In: *Ann. Stat.* 13.1 (1985), pp. 178–203. URL: http://www.jstor.org/stable/2241152.

[122] F. Wang and D. P. Landau. "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States." In: *Phys. Rev. Lett.* 86 (2001), pp. 2050–2053. URL: http://link.aps.org/doi/10.1103/PhysRevLett.86.2050.

[123] H. Wu, A. S. J. S. Mey, E. Rosta, and F. Noé. "Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states." In: *J. Chem. Phys.* 141 (2014), p. 214106. URL: http://scitation.aip.org/content/aip/journal/jcp/141/21/10.1063/1.4902240.

[124] H. Wu and F. Noé. "Optimal estimation of free energies and stationary densities from multiple biased simulations." In: *Multiscale Model. Sim.* 12.1 (2014), pp. 25–54. URL: http://dx.doi.org/10.1137/120895883.

[125] Q. Xia, J. Radzio, K. S. Anderson, and N. Sluis-Cremer. "Probing nonnucleoside inhibitor-induced active-site distortion in HIV-1 reverse transcriptase by transient kinetic analyses." In: *Protein Science* 16.8 (2007), pp. 1728–1737. URL: http://dx.doi.org/10.1110/ps.072829007.

[126] V. M. Zavala, C. D. Laird, and L. T. Biegler. "Interior-point de-
       composition approaches for parallel solution of large-scale non-
       linear parameter estimation problems." In: *Chem. Eng. Sci.* 63.19
       (2008), pp. 4834 –4845. URL: http://www.sciencedirect.com/
       science/article/pii/S0009250907004198.

[127] F. Zhu and G. Hummer. "Convergence and error estimation in
       free energy calculations using the weighted histogram analysis
       method." In: *J. Comput. Chem.* 33.4 (2012), pp. 453–465. URL:
       http://www.ncbi.nlm.nih.gov/pubmed/22109354.

[128] R. Zwanzig. "Nonlinear generalized Langevin equations." In: *J.
       Stat. Phys.* 9.3 (1973), pp. 215–220. URL: http://dx.doi.org/
       10.1007/BF01008729.

ZUSAMMENFASSUNG

Die Dynamik von Proteinen auf langen Zeitskalen ist geprägt durch sogenannte Konformationsänderungen, selten auftretende Veränderungen der räumlichen Struktur von Proteinen. Experimentell können Konformationsänderungen nicht direkt auf atomistischen Skalen beobachtet werden, so dass ein detailliertes Studium dieser Prozesse nur mit Hilfe von Computersimulationen durchgeführt werden kann. Das seltene Auftreten von Konformationsänderungen führt allerdings zu einem Sampling-Problem. Die verlässliche Schätzung von Größen, die mit solchen seltenen Ereignissen in Verbindung stehen, kann nur schwer durchgeführt werden und erfordert zum Teil unvertretbar lange Rechenzeiten für die Erstellung der notwendigen Daten. Markov'sche Modelle erlauben die Schätzung aus einem Ensemble von kurzen Trajektorien, so dass eine einfache Parallelisierung der Berechnung möglich wird. Trotzdem müssen die seltenen Übergänge zwischen verschiedenen Konformationen in den Daten vorhanden sein.

Reversible Markov'sche Modelle können dabei helfen das Sampling-Problem zu lösen. Wahrscheinlichkeiten für seltene Übergänge können mit um Grössenordnungen geringerem Aufwand geschätzt werden, als für die Simulation eines einzelnen Übergangs im Mittel notwendig ist. Diese Modelle ergeben sich natürlicherweise ausgehend von einer diskretisierten Beschreibung der Konformationsdynamik, welche mittels einer endlichen Datenmenge geschätzt wurde.

In der vorliegenden Arbeit werden Methoden für die Schätzung und die Berechnung von Unsicherheiten für reversible Markov'sche Modelle entwickelt. Außerdem wird eine neuartige Methode für die Schätzung von Übergangswahrscheinlichkeiten aus Daten, welche seltene Ereignisse enthalten, vorgestellt. Reversible Schätzer für Übergangswahrscheinlichkeiten werden ausgehend von einer Maximum Likelihood Formulierung konstruiert und effiziente Algorithmen für die Lösung des zugehörigen Optmierungsproblems werden entwickelt. Die vorgestellte Methode kann auch in Situationen angewandt werden in denen zusätzliches Wissen über stationäre Wahrscheinlichkeiten verfügbar ist. Die Berechnung von Unsicherheiten wird durch einen Bayes'schen Ansatz ermöglicht. Ein effizienter Monte Carlo Algorithmus für die Ziehung von reversiblen Übergangsmatritzen aus der zugehörigen Posterior-Verteilung wird entwickelt. Eine Variante bei der das Wissen um stationäre Wahrscheinlichkeiten genutzt werden kann wird ebenfalls vorgestellt. In vielen Fällen ist eine effiziente Schätzung von stationären Wahrscheinlichkeiten aus Enhanced Sampling Simulationen, die nicht unter einem Sampling-Problem leiden, möglich. Häufig ist nur eine Richtung eines Prozesses selten und die entgegengesetzte Richtung kann effizient simuliert werden. Reversibilität kann dazu genutzt werden Übergangswahrscheinlichkeiten auch in solchen Situationen zu schätzen in denen nur eine Richtung des Prozesses beobachtet wurde, wenn Wissen über die stationären Wahrscheinlichkeiten vefügbar ist. Enhanced Sampling Simulationen und kurze Molekulardynamik Trajektorien können so durch reversible Markov'sche Modelle systematisch kombiniert werden.

EIDESSTATTLICHE ERKLÄRUNG

Für die Verfassung der Arbeit wurden keine anderen Hilfsmittel, als die im Text angegebenen, verwendet. Die vorliegende Arbeit ist weder in der vorgelegten noch in einer ähnlichen Fassung zu einem früheren Zeitpunkt in einem Promotionsverfahren eingereicht worden.

Berlin, Juni 2016

Benjamin
Trendelkamp-Schroer