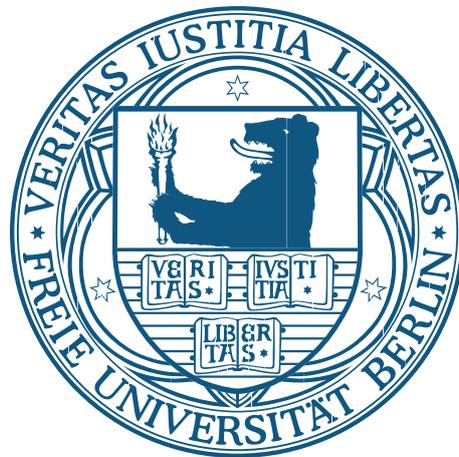


ADVANCED ESTIMATION METHODS FOR  
MARKOV MODELS OF DYNAMICAL SYSTEMS

JAN-HENDRIK PRINZ



Inauguraldissertation  
zur Erlangung des akademischen Grades eines  
Doktors der Naturwissenschaften

vorgelegt beim Fachbereich Mathematik und Informatik  
der Freien Universität Berlin  
von Diplom-Physiker Jan-Hendrik Prinz

Berlin, 2013

Jan-Hendrik Prinz, *Advanced estimation methods for  
Markov models of dynamical systems* © Juli 2012

BETREUER

Dr. Frank Noé  
Freie Universität Berlin  
Fachbereich Mathematik und Informatik  
Arnimallee 2-6  
14195 Berlin

GUTACHTER

Dr. Frank Noé, FU Berlin  
Dr. John D. Chodera, University of California, Berkeley, CA, USA  
Prof. Jeremy C. Smith, Oak Ridge National Lab, Oak Ridge, TN, USA

TAG DER DISPUTATION

7. November 2012

*für meine Großeltern*

2012



---

## ABSTRACT

---

Markov state models (MSM) of molecular kinetics, used to approximate the long-time statistical dynamics of a molecule by a Markov chain on a discrete partition of configuration space, have seen widespread use in recent years. This thesis deals with the improved generation, validation, the application and the extension to experimental observations of these MSM. The four major parts each address different aspects: (1) a summary of the current state of the art in generation and validation of MSMs serving as an introduction along with some important insights into optimal discretization, (2) an investigation of efficient computation and error estimation of the committor, a widely used reaction coordinate, (3) the theory and application on how to generate markov models from multi-ensemble simulations such as parallel tempering using dynamical reweighting, and (4) the extension of MSM theory to non-markovian observations from low-dimensional observed correlations. All parts contain the necessary theory and methods and are applied to artificial and real systems along with an investigation into robustness and error. Thus, this work extends the quality of the computation of key properties and the general construction of markov models for molecular kinetics and allows to alleviate the gap in connecting estimations from simulation and experiment. This is an important step forward toward the long term goal to have the necessary robustness and accuracy for upcoming adaptive MD simulation strategies.



---

## ZUSAMMENFASSUNG

---

In den letzten Jahren haben sich Markov Modelle als effektives und vielseitiges Werkzeug herausgestellt um molekulare Prozesse durch Markov Ketten auf einem diskreten Zustandsraum zu beschreiben und zu analysieren. Die vorliegende Doktorarbeit beschäftigt sich sowohl mit dem Generieren, Validieren, als auch mit den Anwendungen und der Erweiterung auf experimentelle Beobachtungen von Markov Modellen. Dabei werden im Wesentlichen vier Aspekte angesprochen: (1) eine umfassende Zusammenfassung des aktuellen Forschungsstands im Generieren und Validieren von Markov State Modellen, die sowohl als Einleitung dient als auch einige wichtige neue Erkenntnisse zum Problem der Diskretisierung vorstellt, (2) eine effiziente und robuste Berechnung und Fehleranalyse des Kommitors, einer wichtigen Reaktionskoordinate, (3) die Theorie und Anwendung, um mittels Dynamical Reweighting verbesserte Markov Modelle aus einem Satz von Simulationen zu erzeugen, die unter verschiedenen globalen Parametern (z.B. Temperatur) generiert wurden, und (4) die Erweiterung der Markov Modell Theorie auf nicht-markovsche Beobachtungen von experimentellen Trajektorien. Jeder Teil enthält die notwendigen Theorien und Methoden, sowie Anwendungen auf künstliche oder reelle Systeme zusammen mit einer Fehleranalyse. Damit erweitert die vorliegende Arbeit das Feld der Markov Model Theorie um wichtige neue Erkenntnisse mit dem Ziel adaptive molekulardynamische (MD) Simulationen zu ermöglichen und schlägt mittels MSM eine Brücke zwischen MD Simulationen und experimentellen Beobachtungen.



---

## PUBLICATIONS

---

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

- [1] **Prinz, J.-H.**, Wu, H., Sarich, M., Keller, B. G., Senne, M., Held, M., Chodera, J. D., Schütte, C. & Noé, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011). doi [10.1063/1.3565032](https://doi.org/10.1063/1.3565032).
- [2] **Prinz, J.-H.**, Held, M., Smith, J. C. & Noé, F. Efficient Computation, Sensitivity, and Error Analysis of Committor Probabilities for Complex Dynamical Processes. *Multiscale Model. Sim.* **9**, 545–567 (2011). doi [10.1137/100789191](https://doi.org/10.1137/100789191).
- [3] **Prinz, J.-H.**, Chodera, J. D., Pande, V. S., Swope, W. C., Smith, J. C. & Noé, F. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.* **134**, 244108 (2011). doi [10.1063/1.3592153](https://doi.org/10.1063/1.3592153).
- [4] Chodera, J. D., Swope, W. C., Noé, F., **Prinz, J.-H.**, Shirts, M. R. & Pande, V. S. Dynamical reweighting: improved estimates of dynamical properties from simulations at multiple temperatures. *J. Chem. Phys.* **134**, 244107 (2011). doi [10.1063/1.3592152](https://doi.org/10.1063/1.3592152).
- [5] **Prinz, J.-H.**, Keller, B. G. & Noé, F. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.* **13**, 16912–16927 (2011). doi [10.1039/c1cp21258c](https://doi.org/10.1039/c1cp21258c).
- [6] Keller, B. G., **Prinz, J.-H.** & Noé, F. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chemical Physics* **396**, 92–107 (2011). doi [10.1016/j.chemphys.2011.08.021](https://doi.org/10.1016/j.chemphys.2011.08.021).
- [7] Held, M., Metzner, P., **Prinz, J.-H.** & Noé, F. Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys. J.* **100**, 701–710 (2011). doi [10.1016/j.bpj.2010.12.3699](https://doi.org/10.1016/j.bpj.2010.12.3699).
- [8] **Prinz, J.-H.**, Chodera, J. D. & Noé, F. Spectral rate theory for two-state kinetics. *submitted to Physical Review X, ArXiv preprint physics* (2012). url <http://arxiv.org/abs/1207.0225>.



*"Today is a gift – that's why it's called the present!"*

— England, 12. Jh

---

## DANKSAGUNGEN

---

An dieser Stelle möchte ich allen danken, die mich während der Promotion Unterstützt haben und mir so die Vollendung dieser Dissertation ermöglicht haben:

Mein besonderer Dank geht an *Dr. Frank Noé*, zum einen für die Möglichkeit meine Dissertation in seiner Arbeitsgruppe zu schreiben und zum anderen für die sehr persönliche und wissenschaftlich stets produktive und ausführliche Betreuung.

I would also like to thank *Dr. John D. Chodera* for his agreement to be an external corrector and his help with his incredible scientific knowledge and motivating character.

As third corrector and my advisor during my time in Heidelberg I would like to express my gratitude for his support and guidance to *Prof. Jeremy C. Smith*.

Für die vielen fachlichen Gespräche und inhaltlichen Diskussionen und Hilfen möchte mich besonders bei *Bettina Keller, Martin Senne, Marco Sarich* bedanken. Vielen Dank insbesondere an *Prof. Christof Schütte* für seine fachliche Unterstützung und die fürsorgliche Leitung der Biocomputing Arbeitsgruppe der FU Berlin.

*Michelle Krynskaya* sei tausend Dank für die vielen Korrekturen und Hilfe mit der englischen Sprache.

Nicht vergessen möchte ich alle Kollegen, sowohl die der Computational Molecular Biology und Biocomputing Gruppe der Mathematik der FU Berlin als auch die der Computational Molecular Biophysics Group in Heidelberg, die inzwischen ans Oak Ridge National Lab umgezogen ist – vielen Dank für viele wissenschaftlichen Gespräche, offene Ohren bei Problemen und auch mal Zeit für andere Dinge.

Gerne möchte ich mich bei *Martin Held* für die moralische Unterstützung in der Endphase der Arbeit bedanken, sowie bei *Sandra Patzelt-Schütte* und *Martin Senne* für die vielen inspirierenden Diskussionen und aufbauenden Gespräche in unserem Zimmer.

Für die finanzielle Unterstützung danke ich der Universität Heidelberg, der FU Berlin und dem DFG Research Center Matheon.

Vielen Dank an alle Freunde und meine Familie fürs Mitfiebern im Endspurt und das Verständnis für so mache Absage, insbesondere an *Anne Liepelt* für die moralische Unterstützung in den schlechten Zeiten und für das Teilen der Schönen.

Zuletzt möchte ich vor allem meinem Bruder *Marc-André* und meinen Eltern *Christiane* und *Gerd-Willi Prinz* danken, die mich immer und ohne viele Fragen unterstützt haben um mir diese Promotion zu ermöglichen.

Thank you und vielen Dank!

---

## CONTENTS

---

### INTRODUCTION

1	INTRODUCTION	1
2	MARKOV STATE MODELS	7
2.1	Continuous dynamics . . . . .	7
2.2	Transfer operator approach . . . . .	12
2.3	Markov Models vs. Master Equation . . . . .	19
2.4	Discretization . . . . .	19
2.5	Discrete State Space / Transition Matrix . . . . .	31
2.6	Estimation from data and validation . . . . .	34
2.7	Discussion and Conclusion . . . . .	45

### APPLICATION

3	EFFICIENT COMMITTOR COMPUTATION	49
3.1	Introduction . . . . .	49
3.2	Theory . . . . .	50
3.3	Alternative Committed Computation . . . . .	55
3.4	Generalization to multiple states . . . . .	59
3.5	Sensitivity and Uncertainty . . . . .	60
3.6	Applications . . . . .	65
3.7	3D Model . . . . .	72
3.8	Discussion and Conclusion . . . . .	73
4	MULTI-ENSEMBLE ESTIMATION	77
4.1	Introduction . . . . .	77
4.2	Dynamical Reweighting . . . . .	79
4.3	Application to Markov Models . . . . .	83
4.4	Application to alanine dipeptide PT simulation . . . . .	84
4.5	Discussion . . . . .	93
5	ESTIMATION FROM EXPERIMENTAL TIME SERIES	97
5.1	Introduction . . . . .	97
5.2	Classical Approaches . . . . .	98
5.3	Observations and Projections . . . . .	99
5.4	2-State Dynamics . . . . .	107
5.5	Estimation . . . . .	110
5.6	Sensitivity Analysis . . . . .	119
5.7	Spectral Estimation . . . . .	121
5.8	Examples . . . . .	122
5.9	Application to Optical Tweezer Experiments . . . . .	126
5.10	Theoretical RNA model . . . . .	129
5.11	Reconciling force probe experiment and RNA model . . . . .	136
5.12	Summary and Discussion . . . . .	140
6	SUMMARY & OUTLOOK	147

## APPENDIX

A	NOTATION AND SYMBOLS	151
B	MULTI ENSEMBLE ESTIMATION	155
B.1	Efficient solution of the self-consistent equations for canonical distribution of Hamiltonian trajectories . . . . .	155
B.2	Proof that modified PT protocol samples from canonical stationary distribution . . . . .	156
B.3	Convergence of transition probabilities in Bayesian Methods	157
C	MATHEMATICAL DETAILS	159
C.1	Sensitivity Derivations . . . . .	159
C.2	Moore-Penrose pseudoinverse . . . . .	162
C.3	Rate Estimation Error bounds . . . . .	162
D	SYSTEMS SETUP	167
D.1	Exemplary model systems . . . . .	167
D.2	RNA Hairpin . . . . .	169
	BIBLIOGRAPHY	171

---

LIST OF FIGURES

---

Figure 2.1	1D model - Potential and eigensystem decomposition . . . . .	20
Figure 2.2	1D model - Transfer operator . . . . .	21
Figure 2.3	1D model - Timescale Regions . . . . .	22
Figure 2.4	2D model - Eigenvector Approximation . . . . .	28
Figure 2.5	2D model - Chapman-Kolmogorov test . . . . .	46
Figure 3.1	2D model - Energy landscape . . . . .	64
Figure 3.2	2D model - Reference committor . . . . .	64
Figure 3.3	2D model - Comparison of committor predictions . . . . .	66
Figure 3.4	2D model - Comparison of committor variance . . . . .	68
Figure 3.5	2D model - Estimation error predictions . . . . .	69
Figure 3.6	2D model - Total uncertainty prediction . . . . .	69
Figure 3.7	2D model - Comparison of uncertainty contribution . . . . .	70
Figure 3.8	2D model - Net flux . . . . .	71
Figure 3.9	2D model - Multi-state committor . . . . .	71
Figure 3.10	3D model - Equipotential surfaces . . . . .	72
Figure 3.11	3D model - Isocommittor surfaces . . . . .	73
Figure 3.12	3D model - Bundle of association pathways . . . . .	74
Figure 3.13	3D model - Density of reactive trajectories . . . . .	74
Figure 4.1	Distribution of total energies from parallel tempering simulation . . . . .	85
Figure 4.2	Terminally-blocked alanine - structure and potential of mean force . . . . .	86
Figure 4.3	Comparison of all inter-state transition probabilities as a function of temperature . . . . .	88
Figure 4.4	Detailed comparison of transition probability distribution estimates at 302 K . . . . .	90
Figure 4.5	Temperature dependence of estimated eigenvalues . . . . .	91
Figure 4.6	Similarity matrices for different methods at 302 K . . . . .	92
Figure 4.7	Relative contribution to the correlation estimates . . . . .	93
Figure 4.8	Absolute contributions to the estimation of the single correlations . . . . .	94
Figure 5.1	Functional relation between projections and observations . . . . .	101
Figure 5.2	2-state rate estimation example . . . . .	123
Figure 5.3	Comparison of time scale estimation using spectral estimation . . . . .	125
Figure 5.4	Experimental Optical Tweezer Setup . . . . .	128
Figure 5.5	Experimental traces . . . . .	129

Figure 5.6	Estimated Time scales from experimental traces	130
Figure 5.7	Classification of structures of the p21ab RNA Hairpin . . . . .	131
Figure 5.8	RNA Hairpin - Equilibrium Predictions . . . . .	135
Figure 5.9	Schematic analysis of relaxation processes . . . . .	137
Figure 5.10	Approximation of eigenvectors by the theoretical model . . . . .	141
Figure 5.11	Representation of estimated eigenvectors in the theoretical basis . . . . .	142
Figure B.1	Confidence Levels of Transition Matrix Sampling	158

---

## LIST OF TABLES

---

Table 2.1	Matrix Relations used for discrete time, discrete state space MSM . . . . .	33
Table 3.1	List of prior probability distributions . . . . .	66
Table 4.1	List of MSM Estimation Methods . . . . .	84
Table 4.2	Consistency of (PT) trajectories . . . . .	86
Table 4.3	Comparison of error in estimation at 302 K . . . . .	89
Table 5.1	Estimated total trap extensions . . . . .	136
Table 5.2	Definition of the 5 relevant end-to-end distance groups . . . . .	138
Table A.1	Commonly used symbols . . . . .	152
Table A.2	Transition Matrix Notation . . . . .	152
Table A.3	Symbol ornaments . . . . .	153
Table A.4	General Notation . . . . .	153

---

## ACRONYMS

---

dsDNA	double-stranded DNA (Deoxyribonucleic acid)
DR	Dynamical reweighting
GEVP	Generalized Eigenvalue Problem
HMM	Hidden Markov Model
ITS	Implied Time Scale

MCMC	Markov Chain Monte Carlo
MD	Molecular Dynamics
MSM	Markov State Model
MSTIS	Multi-State Transition Interface Sampling
NMR	Nuclear Magnetic Resonance
PCCA	Perron Cluster Cluster Analysis
PMF	Potential of Mean Force
PT	Parallel Tempering
RCQ	Reaction Coordinate Quality
RE	Rate Matrix Estimation
RMSD	Root Mean Squared Displacement
ssRNA	single-stranded RNA (Ribonucleic acid)
SVD	Singular Value Decomposition
TE	Transition Matrix Estimation
TPT	Transition Path Theory



---

## INTRODUCTION

---

Biological macromolecules, like proteins or nucleic acids, are not static structures. They are driven by thermal motion and interactions with their molecular environment while undergoing conformational fluctuations and changing their spatial configurations. These *conformational transitions* are essential to their biological function and span large ranges of length scales, time scales and complexity. This includes folding [9, 10], complex conformational rearrangements between native protein substates [11, 12], and ligand binding [13]. Out of this decades-old proposal that biomolecular kinetics show a complex structure, often involving transitions in a sophisticated network of long-lived, or “metastable” states experiments for a wide variety of biological systems have emerged [14]: With the ever increasing time resolution of ensemble kinetics experiments and the more recent maturation of single-molecule techniques in biophysics, experimental evidence supporting the near-universality of the existence of multiple metastable conformational substates and complex kinetics in biomolecules has continued to accumulate [15, 16, 17, 18, 19, 20, 21]. Enzyme kinetics has been shown to be modulated by interchanging conformational substates [22] and protein folding experiments have found conformational heterogeneity, hidden intermediates, and the existence of parallel pathways [23, 24, 25, 26, 27, 28]. The list of experimental data supporting the existence of such metastable states is long, including the whole range of experimental possibilities, such as Nuclear Magnetic Resonance (NMR) spectroscopy [29, 30, 20], fluorescence emission [31, 21], energy transfer [32, 25], correlation spectroscopy [33, 26], and non-equilibrium perturbation experiments [31]. Aside from a qualitative understanding, the fundamental quest is a quantitative characterization of what gives rise to these essential conformations and their functional interactions. The answers to this will have a significant impact on our understanding of many biological processes, such as, for example, signaling events, enzyme regulation, allostery, and drug design with conformationally flexible molecules.

Although laboratory experiments are designed to resolve kinetic processes and, especially in the case of single-molecule experiments, heterogeneity of some of these processes, the accessible methods are adversely affected by several issues: Firstly, the observations are al-

ways indirect which requires an interpretation of the collected data. Secondly, only spectroscopically resolvable samples can be monitored, and lastly, an acceptable signal-to-noise ratio generally comes at the expense of either time resolution (in single molecule experiments) or the ability to resolve heterogeneity of populations (in ensemble experiments). To fill this gap, Molecular Dynamics (MD) simulations have become a widely accepted tool to investigate structural details of molecular processes in complementary ways and relate them to experimentally resolved features [34, 35, 36].

Traditionally, MD studies often involved a manual analysis based in the subjective interpretation of a few rare events utilizing visualization software and molecular movies. Although visually appealing, these single-molecule analyses may be misleading as they often mask or distort the statistical relevance of such events in the ensemble picture and, conversely, might miss rare but important events altogether. Another common approach, especially in protein folding analyses, is to reduce the dynamics to a projection onto a few user-defined order parameters (e.g. the Root Mean Squared Displacement (RMSD) to a single reference structure, radius of gyration, principal components, or selected distances or angles). The notion of these methods is that carefully chosen order parameters will allow to resolve the slow and relevant kinetics of the molecule while unimportant motions are suppressed. While a simplification of this form directly allows for an intuitive interpretation of the dynamics, these projection techniques have been shown to disguise the true structure of the underlying kinetics. The artificial aggregation of kinetically detached structures and merging of transition and stable conformational regions, might draw an overly simplistic picture of the intrinsically complex kinetics [37, 38, 39].

In order to resolve the important kinetic features such as low-populated intermediates, structurally similar metastable states or structurally distinct – but parallel – pathways, it is inevitable to employ analysis techniques that are sensitive to such details. To generate an analysis understandable to human intuition, clearly, some reduction of high-dimensional biomolecular dynamics data (e.g. from large quantities of MD trajectories) is necessary. Nevertheless, such reductions must be done under mathematical and statistical considerations and have to be guided by the specific structural and kinetic information in this data, rather than by the subjectivity of the analyst. The natural course of action towards modeling the kinetics of molecules is to first partition the conformation space into a discrete set of states [40, 38, 41, 42, 43, 44, 45, 39, 46, 47, 48]. Although this step could still disguise information when lumping states that have an important distinction, it has been proven that a “sufficiently fine” partitioning will be able to resolve a required level of detail [49]. Subsequent to partitioning, transition rates or probabilities between states

can be calculated, either based on rate theories [40, 11, 50], or based on transitions observed in MD trajectories [37, 39, 51, 52, 53, 46, 47]. The resulting models are often called transition networks, Master equation models or Markov State Model (MSM), where “Markovianity” means that the kinetics are modeled by a memoryless jump process between states.

This thesis focuses on “Markov (state) models” (abbreviated here by “MSM” [54]), which model the kinetics with an  $n \times n$  matrix of transition probabilities. Given that the modeled system resides in one of its  $n$  discrete substates, these *transition matrices* contain the conditional (jump) probabilities to find the system in any of these substates a fixed time  $\tau$  later. An essential feature of an MSM is the switching to the ensemble picture of the dynamics instead of the view of the single trajectories which in turn requires the dynamics to be ergodic [55, 56]. Consider an experiment that traces the equilibrium dynamics (i.e. the unperturbed dynamics for a single molecule) of an ensemble of molecules – but starting from a distribution that is out of equilibrium – such as in a laser-induced temperature-jump experiment [57]. Here, the specific sequence of microscopic events occurring during the trajectory of any individual molecule may be of little relevance, as these individual trajectories all differ in microscopic detail and the relevant physical details are statistical properties of the ensemble: time-dependent averages of spectroscopically observable quantities, statistical probabilities quantifying the change in population of conformationally-similar states or the similarity of the dominant trajectory pathways. All of these statistical features can be easily computed from Markov models, as these models already encode the ensemble dynamics [35, 58]. Respectively, individual realizations of almost arbitrary length can be easily generated, simply by generating a random state sequence according to the MSM transition probabilities which is sometimes helpful for the development of human intuition.

Because Markov models recover the *locality* of physics also in the configuration space, i.e. only *conditional* transition probabilities between discretized states are needed to construct a Markov model and no global absolute probabilities, the computational burden can be divided among many simulations using loosely-coupled parallelism, facilitating a “divide and conquer” approach. To estimate the transition probabilities, the single independent trajectories only need to be long enough to reach *local* equilibrium within the discrete state. The time scales are rather short compared to *global* equilibrium relaxation times that may be orders of magnitude longer. In other words, the dependency between simulation length and molecular time scales is largely lost; microsecond- or millisecond-time scale processes can be accurately modeled despite the model having been constructed from trajectories orders of magnitude shorter [35, 59]. Moreover, assessment of the statistical uncertainty of the model can be used to adap-

tively guide the process of model construction, achieving the desired statistical precision with much less computational effort than would be necessary with a single long simulation [60, 61, 35].

Finally, computation of statistical quantities of interest from Markov models is straightforward, and includes:

- Time-independent properties such as the stationary, or equilibrium, probability of states or free energy differences between states [35, 38, 62].
- Relaxation time scales that can be extracted from experimental kinetic measurements using various techniques such as laser-induced temperature jumps, fluorescence correlation spectroscopy, dynamic neutron scattering or NMR [35, 38].
- Relaxation functions that can be measured with non-equilibrium perturbation experiments or correlation functions that can be obtained from fluctuations of single molecule equilibrium experiments [35, 58].
- Transition pathways and their probabilities, e.g. the ensemble of protein folding pathways [35, 63].
- Statistical uncertainties for all observables [64, 58, 61, 60, 2].

Having briefly outlined the importance and possible future impact of Markov models, this thesis deals with several aspects ranging from the construction and usage over the application of MSMs in the context of conformational changes of biomolecules. While previous works have often stressed the model construction from simulations, here we will address the connection to experimental data, trying to make way for more advanced and especially adaptive simulation techniques.

Before continuing with the theory chapter it is advisable to first read to the section on notation and symbols in the appendix A to avoid misunderstandings in the notation and the meaning of frequently used symbols. The thesis is outlined as follows:

Chapter 2 provides a general introduction into the theory, generation and applicability of MSMs in the given biological context and is mainly based on the publication [1]. It is intended to provide the necessary foundations to understand the subsequent chapters and summarizes the current state of the art of theory and methodology for MSMs. In addition to its review-like character, the chapter provides a detailed analysis of the discretization error for MSM and with it important implications on the generation of Markov models based on recently published quantitative upper error bounds. In contrast to previous practice [51, 52, 65, 53] it is shown, that MSMs can be improved if non-metastable states are introduced near the transition states, providing a theoretical basis for the development of efficient adaptive discretization methods.

The three subsequent chapters are each separate projects dealing with connected but not consecutive topics and are based on the MSM theory in chapter 2. Each chapter is self-contained and closed by its own conclusion section:

Chapter 3 is concerned with the efficient and robust calculation of the *committor*, a widely used variable in the analysis of dynamics in terms of MSM. A new viewpoint on the calculation utilizing eigenvectors is presented which allows a physical interpretation in terms of the slow relaxation processes. In addition, the method is extended to a multiple committor with more than two cores that can then be interpreted as a fuzzy decomposition into a discrete number of states. The method is finally applied to several examples of various sizes along a detailed error analysis which provides the necessary foundation for adaptive estimation techniques of the committor.

Chapter 4 is concerned with the construction of MSM in the case where the collected data are chosen from different, but physically related ensembles, e.g. simulations of the same system, but at various ensemble parameters, such as temperature. It uses a new non-parametric method called *Dynamical Reweighting* to combine observations of various related simulations to enhance the statistics at the ensemble parameters of interest. The improvement in statistical precision is demonstrated on data from a previously published parallel tempering simulation which is in particular challenging since the dynamical correlation length in each simulation is much shorter than the longest relaxation time of the system.

Chapter 5 then addresses the issues in the reconstruction of MSM properties from very low dimensional time series, often collected from experimental setups. A new way to circumvent the problem of the projection error in the construction of Markov models from observations is presented by means of relaxation in dealing with the non-Markovianity. This allows to more accurately reconstruct properties from the underlying dynamics which is demonstrated at sample problems and on experimental data collected from an optical tweezer experiment.

Finally, the last chapter 6 will discuss the presented results in a more general context and conclude with an outlook of objectives and challenges for future work. The accompanying appendices contain additional information on the notation (appendix A), supplemental information on the methods and systems used in multi-ensemble estimation chapter (appendix B), extended mathematical proofs (appendix C) and details to the model systems (appendix D).



---

## MARKOV STATE MODELS

---

In the last years Markov models have become a well-established way to describe dynamical processes in biology. In this chapter we want to give a general overview over the standard methods used in the process to generate Markov models. This includes the mathematical basis and justification, the methods involved for estimating the parameters of the Markov model and last the validation against the data for verification. The theory presented in the following is an adaption of the content that was published in the article

- [1] Prinz, J.-H., Wu, H., Sarich, M., Keller, B. G., Senne, M., Held, M., Chodera, J. D., Schütte, C. & Noé, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011). doi [10.1063/1.3565032](https://doi.org/10.1063/1.3565032).

The style of this chapter is intended to resemble a textbook style to support the usage as a reference.

### 2.1 CONTINUOUS DYNAMICS

A variety of simulation models that all yield the same stationary properties, but have different dynamical behaviors, are available to study a given molecular model. Thus, if only stationary properties are of interest, the choice does not matter and one can even change the actual evolution of the system beyond limitations of physical aspects. This is extensively exploited in Monte-Carlo methods to increase the rate of convergence in simulations. We are aiming at the dynamics and therefore the choice of the concrete dynamical model must additionally be guided by a desire to mimic the relevant physics for the system of interest (such as whether the system is allowed to exchange energy with an external heat bath during the course of dynamical evolution), balanced with computational convenience (e.g., the use of a stochastic thermostat in place of explicitly simulating a large external reservoir). Going into the details of these models is beyond the scope of this thesis and therefore we will state the minimal physical properties that we require the dynamical model to obey.

In the following we review the continuous dynamics of a molecular system in thermal equilibrium, and introduce a mathematical object

that characterizes the evolution of this system, the dynamical *propagator*, whose approximation is our primary concern. We now assume that there exists a one-to-one and onto mapping from the unique and instantaneous states of the system onto elements from the mathematical object *state space*  $\Omega$ . Depending on the physical nature of the system, the state space  $\Omega$  is often

continuous, e.g.  $\Omega = \mathbb{R}^n$

or might be

discrete, e.g.  $\Omega \subset \mathbb{N}$

in some special cases (such as spin glasses and on-lattice Go models [66]). The concrete shape of the evolution of a single system can then be described as a mapping from a point in time  $t$  to a unique state  $x \in \Omega$  in the state space which we will denote always by

$$x : t \in \mathbb{R}^+ \mapsto x_t \in \Omega.$$

For most systems, including molecular systems, this means that aside from of the systems spatial coordinates of interest (e.g. the geometrical configuration of a protein) we also have to include the velocities and a description of the environment, e.g. all surrounding heat bath particles. If our model is based upon real-world system, the model should be time-continuous,  $t \in \mathbb{R}^+$  since time itself is continuous. For other models and from an experimental point of view is also reasonable to consider models which are based on discrete timesteps  $\Delta t$  and predict the time-development only at these fixed time-intervals

$$x : k \in \mathbb{N}^0 \mapsto x_k := x_{k\tau} \in \Omega.$$

### 2.1.1 Requirements

The concrete evolution  $x_t$  can be regarded as a particular realization of a stochastic process  $X_t \in \Omega$  with a probability space  $\{\Omega, \mathcal{F}, \mu\}$  over the state space  $\Omega$ , the  $\sigma$ -algebra  $\mathcal{F}$  and a probability measure  $\mu$ . For reasons of simplicity we will not derive the following assumptions and conclusions in the notation of stochastic processes, but rather assume that the probability space can be equipped with the Borel set  $\mathcal{B}(\Omega)$  and a weighted Lebesgue-measure for continuous state spaces  $\Omega$  or a weighted discrete measure for discrete state spaces. In the discrete case we will use probabilities in the original sense while in the continuous case we assume that probabilities of any random variable  $X$  can equivalently be expressed using a continuous probability density function  $f_X$ ,

$$\mathbb{P}(X \in A) = \int_A \mathrm{d}y f_X(y), \forall A \in \mathcal{B}(\Omega)$$

and that this holds for all sets  $A \subset \Omega$  instead of using the  $\sigma$ -algebra. In general, we will assume the following properties to be true for a such a dynamical process  $x_t$  if not stated otherwise:

*A. The dynamics is Markovian*

For the evolution  $x_t$  in its full state space (i.e. phase space) the change of the system, either instantaneous or in non-vanishing time-steps is determined solely by  $x_t$  and no information at any time points  $t_1 < \dots < t_v < t$  is required. This can be formulated as

$$\mathbb{P}[x_{t+\tau} \in A | x_{t_1}, \dots, x_t = x] = \mathbb{P}[x_{t+\tau} \in A | x_t = x], \quad \forall A \in \Omega$$

being true which is called the *Markov property* and hence a Stochastic process with this property is a Markov process.

Further, we assume that the rule of evolution is homogeneous in time, i.e. invariant under a time-shift (which is then referred to as a *time-homogeneous* Markov process). This allows to express the evolution of a particles (and later also of an ensemble of independent particle) at any point in time by either a discrete or a continuous distribution of *jump probabilities*  $p_\tau(x, y)$ . These can be defined by

$$\int_A dx p_\tau(x, y) = \mathbb{P}[x_{t+\tau} \in A | x_t = x], \quad \forall A \subset \Omega. \quad (2.1.1)$$

describing the evolution for a particle at  $x \in \Omega$  to be found a time  $\tau$  later in a set  $A \subset \Omega$ .

The Markov property itself implies the *Chapman-Komogorov-Semi-group property*, which states that the sequential application of evolution has to be in accordance with a single evolution by the total time

$$p_{\tau_1+\tau_2}(x, z) = \int_\Omega dy p_{\tau_1}(x, y) p_{\tau_2}(y, z), \quad \forall z \in \Omega. \quad (2.1.2)$$

The group property is not complete as we do not require the existence of a time-inverse propagation – evolution is always forward in time.

*B. Unique stationary distribution*

We require first that the state space  $\Omega$  cannot be separated into disjoint subsets which are not linked dynamically. This can be expressed mathematically by

$$\forall x, y \in \Omega, \exists t < \infty \text{ s.t. } p_t(x, y) > 0$$

that there is a non-vanishing probability to reach any point  $y \in \Omega$  from any initial point  $x \in \Omega$  within a finite time  $t$ . In this case the Markov process is said to be irreducible. For infinite state spaces we additionally require that all states are positive-recurrent, which means, that the average return time  $r(x)$  of a state  $x \in \Omega$ , given by

$$r(x) = \mathbb{E}(t \cdot p_t(x, x)) = \sum_{k=1}^{\infty} t \cdot p_t(x, x) < \infty$$

is finite. If both holds, then an infinitely long realization  $x_k$  of the Markov process spends a uniquely defined fraction of time

$$\pi(x) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{k=0}^T \delta(x_k - x) \propto \frac{1}{r(x)}$$

in each state  $x \in \Omega$  that is inversely proportional to the average return time  $r(x)$ . This quantity is called the *stationary density*

$$\pi : x \in \Omega \rightarrow \pi(x) \in \mathbb{R}_+$$

or invariant measure  $\mu$  when used as the probability measure instead of the canonical measure.

### C. The dynamics is ergodic

Later we will replace the evolution of single particles by the evolution of ensembles. For this case ergodicity assures that in equilibrium time-averages will converge to the same values as ensemble-averages. This especially means, that an arbitrary distribution of particles will evolve over time to the unique stationary distribution  $\pi$ . For a Markov process it is enough to require a unique stationary distribution and that all states  $x \in \Omega$  are aperiodic.

The stationary density has to correspond to the equilibrium probability inferred from the dynamics (if the dynamics can be approximated by a Markov process!). For dynamics in molecular simulations done for some specific statistical ensemble this should correspond to the equilibrium probability distribution of this statistical ensemble. In particular for the canonical ensemble (NVT) this is the well-known Boltzmann-Distribution. If the Hamiltonian can be separated into a kinetic and a potential part

$$\mathcal{H}(x) = T(x) + U(x)$$

the stationary distribution takes the simple form

$$\pi^{\text{Boltzmann}[\beta]}(x) = Z(\beta)^{-1} \exp(-\beta U(x)). \quad (2.1.3)$$

with a temperature  $T$  depended parameter  $\beta = (k_B T)^{-1}$ ,  $U(x)$  the potential energy of state  $x$  and

$$Z(\beta) = \int_{\Omega} dx \exp(-\beta U(x))$$

being the partition function. See Figure 2.1 for an example.

### D. The dynamics is reversible

In equilibrium the number of particles moving from  $A \subset \Omega$  to  $B \subset \Omega$  are the same as in the opposite direction per time unit. This can be expressed as

$$\pi(x) p_{\tau}(x, y) = \pi(y) p_{\tau}(y, x),$$

which states that the *absolute* probability of a transition from  $x \in \Omega$  to  $y \in \Omega$  is equal for transitions in the opposite direction. This is referred to as *detailed balance* or *microscopic reversibility* and implies that a given trajectory  $x_t$  is equally likely to the trajectory  $x_{[-t]}$  in time-reversed direction if both are weighted with stationary distribution of their initial state. Thus, if started from equilibrium, from the trajectory itself, one cannot deduce a direction in time.

This is different from stating that the time-reversed ensemble dynamics fulfills the same equations of motion, i.e. that the evolution of an ensemble of particles is symmetric under the change of sign in time which is often referred to as *reversible dynamics*. Here, we require only the reversibility for a trajectory of a single particle, but not for ensembles. For example, Brownian dynamics fulfills our requirement of (microscopic) reversibility, but it is not time-reversible for ensembles since any initial distribution diverges over time irrespective of propagating forward or backwards in contrast to e.g. Hamiltonian dynamics. The connecting property is determinism: If a dynamics fulfills detailed balance and is deterministic it is also reversible in the macroscopic sense.

Although detailed balance is not necessary to construct a Markov model, it is reasonable from a physical point of view: For a system which is in equilibrium and does not obey detailed balance there exists a closed circle of states, that has a higher probability to be traversed in one direction compared to the opposite one. Exploiting this, one could build a device that could extract work from this imbalance. This would contradict the second law of thermodynamics and conversely implies, that there is some other mechanism than thermal energy that drives the system. We will exclude these cases and state, that *in equilibrium* means in thermal equilibrium and thus our dynamics must fulfill detailed balance in order to not be in violation of the laws of thermodynamics.

### 2.1.2 Applicability

After stating the requirements, the question remains whether these can be fulfilled. Data measured directly from nature using experiments will preserve the Markov property unless the experimental setup interferes, which is an important aspect, but has to be addressed by the experimentalist. Mostly, we use time series from computer simulations where the used integrator and, for protein simulation, additional numerical methods to ensure ensemble constraints might include systematic errors. In general, we can state that the above conditions do not place overly burdensome restrictions on the choices of dynamical models used to describe at least equilibrium dynamics. Most stochastic thermostats are consistent with the above assumptions, e.g. Andersen [67] (which can be employed with either

massive or per-particle collisions, or coupled to only a subset of degrees of freedom), Hybrid Monte Carlo [67], over-damped Langevin (also called Brownian or Smoluchowski) dynamics [68, 69], and step-wise-thermalized Hamiltonian dynamics [53]. When simulating solvated systems, a weak friction or collision rate can be used; this can often be selected in a manner that is physically motivated by the heat conductivity of the material of interest and the system size [67].

While, technically speaking, a Markov model analysis can be constructed for any choice of dynamical model, it must be noted that several popular dynamical schemes violate the assumptions above. Using them means that one is (currently) doing so without a solid theoretical basis, such as regarding the boundedness of the discretization error analyzed in section 2.4. Counter examples include e.g. Nosé-Hoover and Berendsen thermostats that are either not ergodic or do not generate the correct stationary distribution for the desired ensemble [70]. Also, energy-conserving Hamiltonian dynamics, even when considering a set of trajectories that are in initial contact with a heat bath, is not ergodic and therefore ensembles might not converge to the stationary distribution.

We note that the use of finite-time step integrators for these models of dynamics can sometimes be problematic, as the phase or configurational space density sampled can differ from the density desired. Generally, integrators based on symplectic Hamiltonian integrators (such as velocity Verlet [71]) offer greater stability for our purposes.

## 2.2 TRANSFER OPERATOR APPROACH

Instead of giving a probabilistic description of the evolution of a single particle, we can also track the evolution of an ensemble of particles which we already addressed shortly in the discussion about ergodicity. We assume that our ensemble of particles consist of independent realizations of the same dynamics, i.e. they do not interact with each other and they are distributed according to a probability density function

$$p : x \in \Omega \mapsto p(x) \in \mathbb{R}_0^+, \quad \int_{\Omega} dx p(x) = 1.$$

Positivity and the normalization condition assures that

$$p \in L_{\pi}^2 = \left\{ v \mid v : \Omega \mapsto \mathbb{R} \mid \int_{\Omega} dx \pi(x) v(x)^2 < \infty \right\}$$

holds and the time-evolution of a probability function can be expressed as a propagation in a  $\pi$ -weighted  $L_{\pi}^2$ -space<sup>1</sup>. To access all the

<sup>1</sup> The more general approach is to use the stationary distribution as the measure in the  $L_{\pi}^2$ -space, which is, in the cases we treat here, equivalent to a weighted  $L_{\pi}^2$ -space with the canonical measure.

features a Hilbert-Space provides we chose to use the natural scalar product induced by the stationary distribution

$$\langle a, b \rangle = \int_{\Omega} dx \pi(x) a(x) b(x)$$

and with it, the induced norm

$$\|a\| = \sqrt{\langle a, a \rangle}.$$

If not stated otherwise we will *always* use the induced ( $\pi$ -weighted) scalar product. As we will see later, this choice is convenient since the main object of our concern, the transfer operator, will (usually) be self-adjoint w.r.t. this scalar product and thus its eigenfunctions are orthogonal under this scalar product.

It is now reasonable to define a reweighting operator  $\Pi$

$$\Pi : L_{\pi}^2 \rightarrow L_{\pi}^2$$

$$[\Pi p](x) = \pi(x) p(x)$$

to switch between elements of  $L_{\pi}^2$  and the isomorphic dual space which will prove to be useful later. Within this Hilbert-Space we can define the propagator  $Q_{\tau}$  that will evolve a probability distribution  $p_t$  given at time  $t$  into a distribution  $p_{t+\tau}$  a lag time  $\tau$  later in accordance with the single particle evolution given by the jump probabilities  $p_{\tau}(x, y)$ . The acting of the propagator can be written as

$$p_{t+\tau}(x) = [Q_{\tau} p_t](x) = \int_{\Omega} dy p_{\tau}(y, x) p_t(y) \quad (2.2.1)$$

or in short

$$p_{t+\tau} = Q_{\tau} p_t.$$

If we do not talk about lag time  $\tau$  dependence or if it is not stated otherwise, we will omit the indication of the lag time  $\tau$  and use e.g.  $Q := Q_{\tau}$ . For purposes of mathematical simplicity we also introduce an equivalent description in form of the *transfer operator*

$$T_{\tau} : L_{\pi}^2 \rightarrow L_{\pi}^2$$

[55, 72], which acts on functions

$$v_t = \Pi^{-1} p_t$$

that are simply probability distributions reweighted with the inverse of the stationary distribution.

These functions can be considered as isomorphic representations of elements in the dual space. Using the reweighting operator  $\Pi$  we can simply write the transfer operator in terms of the propagator as

$$T_{\tau} = \Pi^{-1} Q_{\tau} \Pi \quad (2.2.2)$$

and we should note, that the transfer operator is *not* the adjoint propagator, which we will introduce later. The propagator  $Q$  and the transfer operator  $T$  are essentially the same object only in two different representations chosen w.r.t. to the invariant measure in  $L^2_\pi$ .

If the Markov process fulfills the requirements from section 2.1.1 (i.e. ergodicity) the dynamics will relax *any* given initial density  $p_t$

$$\lim_{\tau \rightarrow \infty} p_{t+\tau} = \lim_{\tau \rightarrow \infty} Q_\tau p_t = \pi$$

to its unique stationary distribution  $\pi$ .

Although the definition of the propagator  $Q$  or transfer operator  $T$  in Eq. (2.2.1) is formal, it can be given in a concrete form, depending on the kind of dynamics [55]. Independent from this concrete form it has the following properties which are based on the Hilbert-Space  $L^2_\pi$  and the properties of the Markov Process:

#### A. Eigenvalues and Eigenvectors

The propagator  $Q_\tau$  has a set of eigenfunctions  $\phi_i$  and associated eigenvalues  $\lambda_i$  with  $i \in I$  in some (usually infinite but countable) index set  $I$  (see Figure 2.1 and Figure 2.3)

$$Q \phi_i = \lambda_i \phi_i \quad (2.2.3)$$

and the definition of  $T_\tau$  implies that the eigenvectors  $\psi_i$  of  $T_\tau$  are related by

$$\phi_i = \Pi \psi_i.$$

From the Perron-Frobenius theorem it follows that both, the propagator and the transfer operator, have a spectral radius of exactly one, i.e. their eigenvalues lie in the unit circle in the complex plane  $|\lambda_i| \leq 1$ . Ergodicity finally assures, that there is only one unique eigenvalue/eigenvector pair to the eigenvalue of one  $\lambda_1 = 1$  that has the greatest norm [55]. The first eigenfunctions  $\phi_1$  and  $\psi_1$  take special forms:

$$Q_\tau \pi = 1 \cdot \pi = \phi_1,$$

the stationary distribution  $\pi$  and

$$T_\tau \mathbf{1} = 1 \cdot \mathbf{1} = \psi_1$$

the constant function  $\mathbf{1}$  (see Figure 2.1c).

#### B. Reversibility

In the context of reversibility we first introduce the so-called backward propagator  $Q^-$ . To explain its meaning, we go back to the picture of single particle dynamics in equilibrium<sup>2</sup> and ask for the

<sup>2</sup> *In equilibrium* means here, that we look at a particle which is chosen randomly from an ensemble which is in equilibrium and where no driving forces are present.

transition probabilities for ensembles under the time-reversed process, i.e. a process where the probabilities to observe a particular time series has been replaced by the probabilities of the same trajectory observed in time-reversed order. If for a forward time series  $x_t$  the *absolute* probabilities for a transition from  $x \in \Omega$  to  $y \in \Omega$  were given by  $\pi(x) p(x, y)$ , then, for the backward/time-reversed propagation we have to choose the absolute probability to come from  $y$  and end in  $x$  respectively given by  $\pi(y) p_\tau(y, x)$ . The *conditional* transition probability to come from  $y$  when being in  $x$  is then constituting the backward transition probabilities  $p_\tau^\dagger(x, y)$  given by

$$p_\tau^\dagger(x, y) \equiv \frac{\pi(y)}{\pi(x)} p_\tau(y, x)$$

representing the well-defined backward propagator  $Q_\tau^-$ . If we define a *transposed* propagator  $Q^\top$  using

$$[Q_\tau^\top p](x) \equiv \int_\Omega dy p_\tau(x, y) p_t(y) \quad (2.2.4)$$

the backward propagator can be expressed by

$$Q_\tau^- \equiv \Pi Q_\tau^\top \Pi^{-1}$$

In the case of the backward transfer operator we get

$$T_\tau^- \equiv \Pi^{-1} T_\tau^\top \Pi$$

and we can express this relationship using the induced ( $\pi$ -weighted) scalar product as

$$\langle T_\tau^\dagger f, g \rangle \equiv \langle f, T g \rangle, \quad \forall f, g \in L_\pi^2 \quad (2.2.5)$$

and see that the backward transfer operator is the adjoint of the transfer operator w.r.t. the scalar product induced by the stationary distribution. Interestingly, if detailed balance

$$\pi(x) p_\tau(x, y) = \pi(y) p_\tau(y, x)$$

$$\Pi Q^\top = Q \Pi$$

holds, then forward and backward propagator  $Q_\tau^- = Q_\tau$  coincide, and also forward and backward transfer operator,  $T_\tau^- = T_\tau$ . From Eq. 2.2.5 then follows that the transfer operator  $T_\tau$  is self-adjoint w.r.t. the induced scalar product, while the propagator is self-adjoint w.r.t. a scalar product that uses the inverse of the stationary distribution  $\pi^{-1}$ . The backward propagator and transfer operator will be used again in chapter 3.

Conclusively, since the transfer operator is bounded and self-adjoint (if detailed balance holds), there exists a (usual infinite but countable)

set of eigenfunctions  $\psi_i$  of the transfer operator  $T_\tau$  that are orthogonal w.r.t. the weighted scalar product

$$\lambda_i \neq \lambda_j \Rightarrow \langle \psi_i, \psi_j \rangle = 0, \quad \forall i \neq j$$

provided that the corresponding eigenvalues are different. Self-adjointness also implies that all eigenvalues  $\lambda_i$  of  $T_\tau$  have to be real-valued and hence  $\lambda_i \in (-1, 1]$ ,  $i \in I$ .

We can now use the *Chapman-Komogorov-Semi-group property* in Eq. 2.1.2 to infer an important lag time  $\tau$  dependence of the eigendecomposition. In terms of operators it simply translates into

$$Q_{\tau_2+\tau_1} = Q_{\tau_2}Q_{\tau_1}$$

which implies that the eigenfunctions  $\phi_i$  of the propagator  $Q_\tau$  are independent of the lag time  $\tau$  while the eigenvalues are related by

$$\lambda_i(k \cdot \tau) = \lambda_i^k(\tau), \quad k \in \mathbb{Q}^+, \tau > 0.$$

Finally, we get the important relation, that in case of detailed balance, the eigenvectors  $\psi_i$  of the transition operator  $T_\tau$  equal the eigenvectors of the adjoint operator  $Q_\tau^T$  which is in the discrete case often used to identify the *left* eigenvectors of the transition matrix with the eigenfunctions  $\phi_i$  of the propagator  $Q$ .

Lastly, we note a few technical details: We agree on the normalization

$$\langle \psi_i, \psi_i \rangle = 1$$

which will result in

$$\begin{aligned} \psi_1(x) &= 1 \\ \int_{\Omega} \mathbf{d}x \phi_1(x) &= \int_{\Omega} \mathbf{d}x \pi(x) = 1. \end{aligned}$$

Even though the eigenspectrum  $\{\lambda_i \mid i \in I\}$  is usually continuous we only distinguish a finite number of  $m$  dominant eigenvector/eigenvalue pairs with the largest absolute value. By convention we sort all pairs descending by the absolute values of their eigenvalue

$$\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_m|$$

and consider the reminder of the spectrum confined to a ball of radius  $r \leq |\lambda_m|$  centered on 0 in the complex plane. Even if detailed balance does not hold, it can be shown, that if the dynamics is reversible *enough*, the dominant eigenvalues are real-valued and that the complex-valued eigenvalues can be contained in ball in the complex plane with a radius strictly smaller than one [55]. In the following we will only deal with the transfer operator since it carries the same information as the propagator and allows for an easier notation and the usage of the natural scalar product.

### 2.2.1 Timescale Separation

Using the eigenvector/eigenvalue pairs  $\{\psi_i, \lambda_i\}$  the transfer operator  $T_\tau$  can be expanded into a sum of operators as

$$T_\tau = \sum_{i \in I} \lambda_i(\tau) X(\psi_i) \quad (2.2.6)$$

with the projection operator

$$X(v) = v \frac{\langle v, \cdot \rangle}{\langle v, v \rangle}$$

that projects any function onto the subspace spanned by  $v$ . This idea permits an illustrative physical interpretation: In general, we can expand any probability distribution  $p$  into a basis spanned by the eigenvectors  $\psi_i$ . In this basis the propagation is simply an exponential decay, except for the stationary distribution with  $\lambda_1 = 1$  which is always present. Therefore, the dynamics can be regarded as single processes that push the distribution  $p$  back to equilibrium  $\pi$  each with a distinctive speed given by the eigenvalues  $\lambda_i$ . The processes are then indicated by the eigenvectors  $\psi_i$  since these determine the basis transformation, while the  $\lambda_i$  for  $i \in \{2, \dots, m\}$  correspond to a physical time scale

$$t_i = -\frac{\tau}{\ln \lambda_i}, \quad (2.2.7)$$

which is often called the  $i$ -th Implied Time Scale (ITS) [53]. In the limit of long time steps,  $k \rightarrow \infty$ , only the stationary distribution survives since

$$\lim_{k \rightarrow \infty} \lambda_i^k \rightarrow 0, \quad i \neq 1.$$

We now separate the eigenvalues into 3 subsets, that then constitute one part each of the total dynamics:

1. The *stationary* part for  $\lambda_1 = 1$  which is turned into

$$T^{\text{stat}} v = \mathbf{1}$$

2.  $(m - 1)$  *dominant* (slow) real eigenvalues with  $\lambda_i \geq \lambda_m$  which give

$$T_{k\tau}^{\text{slow}} = \sum_{i \in \{2, \dots, m\}} \lambda_{ii}^k X(\psi_i)$$

and will be later be identified with the dominant dynamical part we are interested in

3. and the remaining *fast* contributions with  $|\lambda_i| < |\lambda_{m+1}|$  into

$$T^{\text{fast}}(\tau),$$

which we want to neglect if the time scale  $\tau$  is large enough.

This means for the propagation of  $k$  equidistant timesteps of length  $\tau$  that the transport operator can be written as

$$\begin{aligned} T_{t+k\tau} &= T^{\text{stat}} + T_{k\tau}^{\text{slow}} + T_{k\tau}^{\text{fast}} \\ &= \mathbf{1}\langle \mathbf{1}, \cdot \rangle + \sum_{i=2}^m \lambda_i^k X(\psi_i) + T_{k\tau}^{\text{fast}}. \end{aligned} \quad (2.2.8)$$

This decomposition requires all three subspaces  $T^{\text{stat}}$ ,  $T^{\text{slow}}$  and  $T^{\text{fast}}$  to be orthogonal, which is a direct consequence of detailed balance.

When observing the expansion into eigenfunctions (Eq. 2.2.8), the dependence of the time step  $t + k\tau$  is left only in the exponential of  $\lambda_i^k$  so that with increasing number of timesteps  $k$  or longer lag times  $\tau$ , the number of contributing processes decreases. This allows an approximation of the dynamics (see Figure 2.3) in various time regions. Especially if the lag time  $\tau$  exceeds a certain minimal lag time  $\tau > \tau^{\text{fast}}$ , only the dominant processes are present. We choose

$$\tau^{\text{fast}} = \frac{\log \epsilon}{\log \lambda_{[m+1]}}$$

where  $\epsilon < 10^{-4}$  is chosen *small* enough to cancel the influence of  $T^{\text{fast}}$  to the dynamics and we can set approximately  $T^{\text{fast}} \approx 0$ . Note, that if the spectral gap between the fastest dominant process  $\lambda_m$  and the next slowest processes  $\lambda_{[m+1]}$  is small, then in this choice of  $\tau^{\text{fast}}$  also the fastest dominant process (and even more) vanishes which is usually not desired. We will assume that in the analyzed systems we can (at least to some degree) separate fast and slow processes by choosing an appropriate number  $m$  of dominant processes and a suitable separation lag time  $\tau^{\text{fast}}$ . Even if this is not the case the presented methods are still valid, only their convergence behaviour will be worse.

In Figure 2.1, the second process,  $\psi_2$ , corresponds to the slow ( $\lambda_2 = 0.9944$ ) exchange between basins A+B and basins C+D, as reflected by the opposite signs of the elements of  $\psi_2$  in these regions (Figure 2.1c). The next-slowest processes are the A $\leftrightarrow$ B transition and then the C $\leftrightarrow$ D transition, while the subsequent eigenvalues are clearly separated from the dominant spectrum and correspond to much faster local diffusion processes. The three slowest processes effectively partition the dynamics into four metastable states corresponding to basins A, B, C and D, which are indicated by the different sign structures of the eigenfunctions (Figure 2.1c). The metastable states can be calculated from the eigenfunction structure, e.g. using the Perron Cluster Cluster Analysis (PCCA) method[43, 51]. Figure 2.1d shows the projected eigenfunctions of the transfer operator,  $X^\perp \psi_i$ , onto the crisp subsets indicated by the states A, B, C and D. While the projection error is small for the projections of the four dominant processes the two next

processes almost vanish, indicating that these processes will not be present in dynamics projected onto these 4 states.

### 2.3 MARKOV MODELS VS. MASTER EQUATION

Alternatively to  $Q_\tau$  and  $T_\tau$  which describe the transport of densities exactly by a chosen time-discretization  $\tau$ , one could investigate the density transport with a time-continuous operator  $L$  called *generator*, which is the basis of rate matrices, frequently used in physical chemistry [44, 73] and being related to the Fokker-Planck equation [74].

The generator can be defined as an infinitesimal small transport operator by

$$L = \lim_{\tau \rightarrow 0} \frac{1}{\tau} (Q_\tau - \text{Id}) \quad (2.3.1)$$

if it exists. This allows to express the time evolution of an ensemble  $p$  by a first-order differential equation

$$\partial_t p_t = L p_t.$$

It has to be noted that one can always construct a propagator  $Q_\tau$  for a lag time  $\tau$  from a generator using

$$T_\tau = \exp(\tau L) \quad (2.3.2)$$

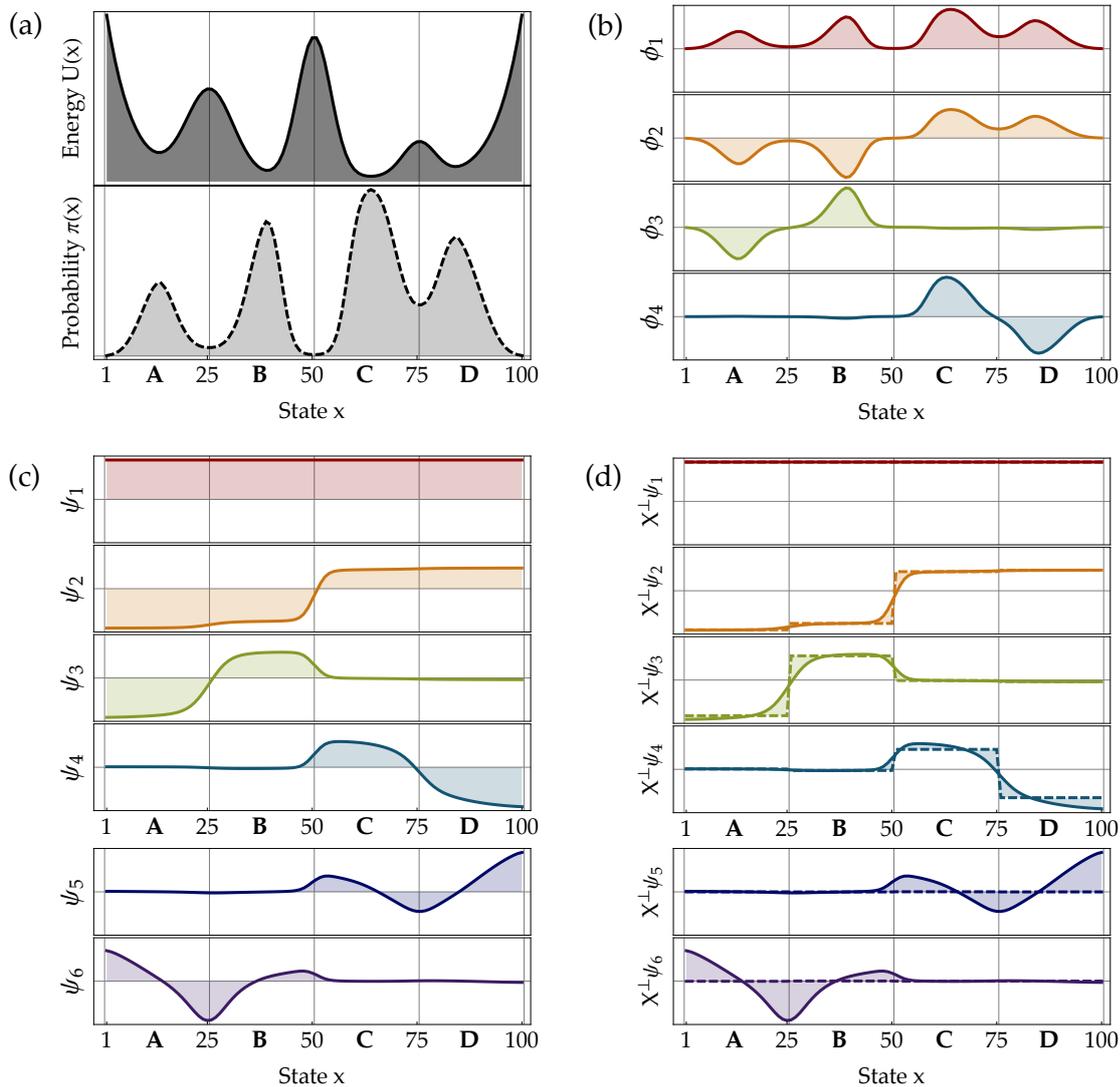
where we used the Taylor expansion of the exponential function to define the exponential of an operator by

$$\exp(\tau L) = \sum_{k=0}^{\infty} \frac{1}{k!} \tau^k L^k.$$

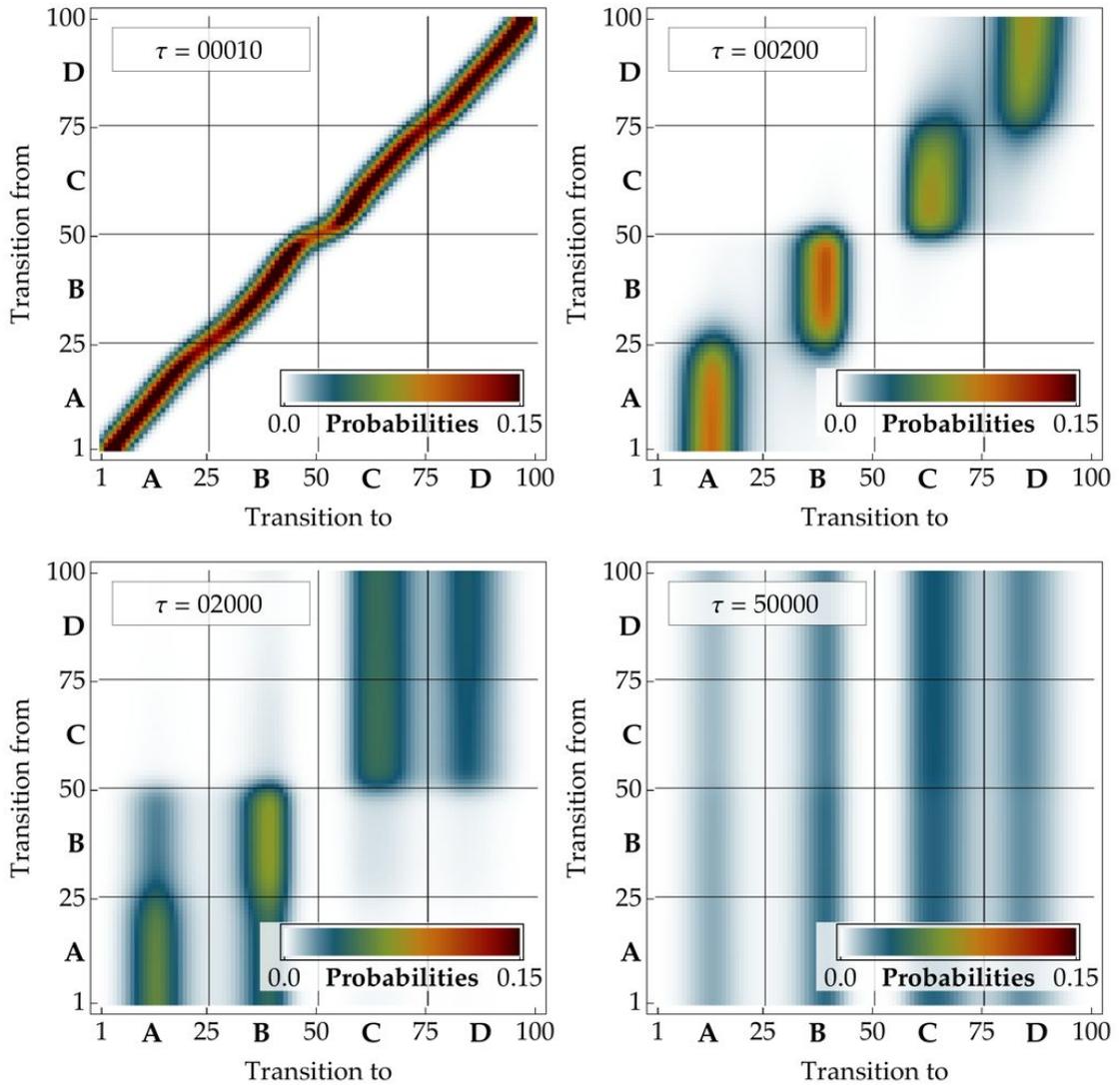
From Eq. (2.3.2) then follows, that the generator and propagator share the same set of eigenfunctions and that the eigenvalues  $\mu_i$  of the generator are given by  $\mu_i = \ln(\lambda)$ . This explains that the generator only uniquely exists, if all eigenvalues of the propagator are real valued (which is true when detailed balance holds) and are also positive,  $\lambda_i \in \mathbb{R}^+$ ,  $\forall i \in I$ . Conclusively, the generator has one single zero eigenvalue  $\mu_1 = 0$  and all other eigenvalues are negative  $\mu_i < 0, \forall i \neq 1$ . In the following we focus on using the transfer operator description.

### 2.4 DISCRETIZATION

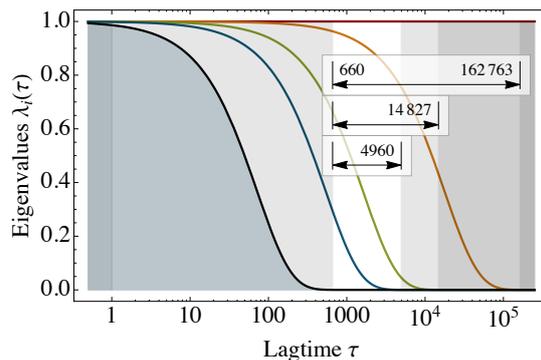
In the context of Markov models we cannot neglect that, for computational purposes, we have to discretize the original full Markov model. In principle, there would be no objections to reduce the problem to



**Figure 2.1 – 1D model potential and eigensystem decomposition** (a) Potential energy function with four metastable states and corresponding stationary density  $\pi(x)$ , (b) The four dominant eigenfunctions  $\phi_1, \dots, \phi_4$  of the transfer operator  $T$  weighted with the stationary density  $\pi(x)$  which correspond to the eigenfunctions of the propagator, (c) The four dominant eigenfunctions of the transfer operator,  $\psi_1, \dots, \psi_4$ , which indicate the associated dynamical processes, and the next two slowest eigenfunctions  $\psi_5, \psi_6$ . The first eigenfunction is associated to the stationary process, the second to a transition between  $A + B \leftrightarrow C + D$  and the third and fourth eigenfunction to transitions between  $A \leftrightarrow B$  and  $C \leftrightarrow D$ , respectively. The two next eigenfunctions correspond to processes within the metastable states, (d) The four dominant and the two next eigenfunctions of the transfer operator projected onto the 4-states  $A, B, C, D$  given by  $X^\perp \psi_1, \dots, X^\perp \psi_6$ . While the projection is quite good for the four dominant eigenfunctions and thus a small projection error, the projection of the two next eigenfunctions almost vanishes, indicating that these processes will not be present in the projected dynamics.



**Figure 2.2 – Transfer operator** Density plot of the transfer operator for the simple diffusion-in-potential dynamics defined on the range  $\Omega = \{1, \dots, 100\}$  (see Appendix D) for lag times  $\tau = \{10, 200, 2000, 50000\}$ . Red indicates high transition probability, white zero transition probability. Of particular interest is the nearly block-diagonal structure where the transition density is large within blocks allowing rapid transitions within metastable basins, and small or nearly zero for jumps between different metastable basins. Depending on the lag time  $\tau$  the number of metastable sets changes.



**Figure 2.3 – Timescale Regions** The development of eigenvalues  $\lambda_i(\tau)$  vs lag time  $\tau$ . Red, yellow, green and blue correspond to the four dominant processes. Grey bars on the right indicate the regions, where the dominant processes disappear while grey bars on the left indicate region where the fast processes are present and at the very left the region not accessible by sampling due to a finite sampling rate. The arrows indicate the range of lag times where the 3 dominant processes are observable. The white region in the middle is the region, where all four dominant processes are exclusively present and thus the approximation by the dominant spectral part is good. The gap between the four metastable processes ( $\lambda_i \approx 1$ ) and the fast processes (blue region) is clearly visible.

one with a finite (and small) state space, but the observed projected dynamics is in almost all cases not Markovian anymore. That means, that any parametrized Markov model based on this observation cannot or only approximately reflect the real dynamics, which is what we are aiming for when building a model for the system in any simulation or experiment. In this section we will deal with the problem of discretization and errors involving the construction of a Markov model from it. In general, we have to distinguish three types of errors:

1. The *spectral error* from neglecting the fast part of the dynamics,
2. the *projection or discretization error* arising from a loss in information by the projection onto a finite-dimensional subspace, and
3. the *statistical error* caused by estimations based on observations of finite length (insufficient data)

where, in this section, we are mainly dealing with the first two systematic errors.

In practical use, the Markov model is not obtained by actually discretizing the continuous propagator although the dynamics is still based on it. Instead, one defines a discretization of state space and then estimates the corresponding discretized transfer operator from a finite quantity of simulation data, such as several long or many short Molecular Dynamics (MD) trajectories that transition between these

discrete states. The statistical error in estimating the parameters of a finite state Markov model from an also finite set of data is dealt with later.

While molecular dynamics in full continuous state space  $\Omega$  is Markovian by construction, the term Markov *model* is due to the fact that in practice, state space must be somehow discretized in order to obtain a computationally tractable description of the dynamics. The Markov model then consists of the partitioning of state space used together with the transition matrix modeling the jump process of the observed trajectory projected onto these discrete states. However, this jump process is no longer Markovian, as the information where the continuous process would be within the local discrete state is lost in the course of discretization. Modeling the long-time statistics of this jump process with a Markov process is an approximation, i.e., it involves a *discretization error*.

This error is a *systematic error*, since it causes a deterministic deviation of the Markov model dynamics from the true dynamics that persists even when the statistical error is excluded by excessive sampling. In order to focus on this effect alone, it is assumed in this section that the statistical estimation error is zero, i.e., transition probabilities between discrete states can be calculated exactly. The results suggest that the discretization error of a Markov model can be made small enough for the Markov State Model (MSM) to be useful in accurately describing the relaxation kinetics, even for very large and complex molecular systems.

#### 2.4.1 Discretization of state space

The most prominent case is the clustering of parts of the state space into a finite number of macro states which will then define a so called crisp clustering. For a crisp partitioning or clustering we define a countable set  $M$  of subsets of the full state space  $\omega_i \subset \Omega$ ,  $i \in M$ , called *macro states*, which are mutually exclusive states  $\omega_i \cap \omega_j = \emptyset$ ,  $i \neq j$  and have associated indicator functions

$$\chi : i \in M, x \in \Omega \mapsto \chi_i(x) = \mathbf{1}_{\omega_i}(x) \in \{0, 1\} \quad (2.4.1)$$

that measure if a certain state  $x$  belongs to macro state  $i$ . More general, we can define *membership functions* that allow the splitting of the probability between several macro states. Similar to the crisp case, this is a generalization and fulfills the same requirements as a partition of unity with strict non-negative functions

$$\chi : i \in M, x \in \Omega \mapsto \chi_i(x) \in [0, 1] \quad (2.4.2)$$

and

$$\sum_{i \in M} \chi_i(x) = 1, \forall x \in \Omega.$$

For a clustering, be it fuzzy or crisp, we can define a projection operator  $X$ , that projects any function  $p(x)$  onto the subspace induced by the clustering  $\chi$ . Using a symmetric mass matrix

$$\mathbf{M} : M \times M \mapsto M_{ij} \in [0, 1] \subset \mathbb{R}^+$$

defined by

$$\mathbf{M}_{ij} = \langle \chi_i, \chi_j \rangle$$

that measures the overlap of two membership functions, the projection operator takes the form

$$X = \sum_{i,j \in M} (\mathbf{M}_{ij})^{-1} \chi_i \langle \chi_j, \cdot \rangle \quad (2.4.3)$$

and the orthogonal projection is given by

$$X^\perp \equiv \text{Id} - X$$

If the clustering is crisp, the mass matrix is diagonal and the projection can be reduced to a superposition of single projections

$$X \equiv \sum_{i \in M} \langle \chi_i, \chi_i \rangle^{-1} \chi_i \langle \chi_i, \cdot \rangle.$$

#### 2.4.2 Quantifying the discretization error

The unavoidable discretization leads to a quantitative analysis of the systematical error induced by the projection. In Markov models of molecular dynamics, this state space reduction usually consists of both, a neglect of degrees of freedom and an additionally discretization of the remaining ones. Formally, all of these operations aggregate sets of points in the continuous state space  $\Omega$  into discrete macro states, and the question to be addressed is what is the magnitude error caused by treating the non-Markovian projected jump process between these sets as a Markov chain. We will deal with an alternative view on the projection error in chapter 5.

The projected transfer operator  $T_\tau^X$  that propagates in subspace spanned by the projection operator  $X$  is defined by

$$T_\tau^X = X T_\tau X, \quad (2.4.4)$$

first projecting, then a transport using the original transfer operator and afterwards again a projection. The projected operator can now be used to propagate any distribution spanned by the membership functions  $\chi$ , i.e. it is closed w.r.t. to  $X$ . This definition is unique up to the specification of the lag time  $\tau$  used for the parametrization and causes a certain ambiguity in the propagation: To propagate an arbitrary distribution in the projected subspace  $X p_t$  for  $k$  timesteps

of length  $\tau$  we can either use the original (exact) propagation and project afterwards

$$\begin{aligned} \mathbf{p}_{t+k\tau}^{\text{orig}} &= \mathbb{X} \mathbf{T}_{k\tau} \mathbb{X} \mathbf{p}_t \\ &= \mathbb{X} \mathbf{T}_\tau^k \mathbb{X} \mathbf{p}_t \end{aligned}$$

or we directly use the projected and thus approximated transfer operator  $\mathbf{T}_{k\tau}^{\mathbb{X}}$

$$\begin{aligned} \mathbf{p}_{t+k\tau}^{\text{proj}} &= \mathbb{X} \mathbf{T}_{k\tau}^{\mathbb{X}} \mathbb{X} \mathbf{p}_t \\ &= \mathbb{X} (\mathbb{X} \mathbf{T}_\tau \mathbb{X})^k \mathbb{X} \mathbf{p}_t \\ &= \mathbb{X} (\mathbf{T}_\tau \mathbb{X})^k \mathbf{p}_t \end{aligned}$$

where in the last step the idempotency

$$\mathbb{X} \mathbb{X} = \mathbb{X}$$

of the projection operator  $\mathbb{X}$  was used. Due to the intermediate projections in the second case, both solutions differ in most cases, but we can quantify the approximation error

$$\begin{aligned} \epsilon(k) &= \left\| \mathbf{p}_{t+k\tau}^{\text{orig}} - \mathbf{p}_{t+k\tau}^{\text{proj}} \right\| \\ &= \left\| (\mathbb{X} \mathbf{T}_\tau^k \mathbb{X} - \mathbb{X} [\mathbf{T}_\tau \mathbb{X}]^k) \mathbf{p}_t \right\| \end{aligned}$$

by measuring the difference between both solutions as a function of the number of time steps  $k$  used. To proceed we define the *eigenfunction approximation error*

$$\delta_i := \|\psi_i - \mathbb{X} \psi_i\| = \left\| \mathbb{X}^\perp \psi_i \right\|, \quad i \in \{1, \dots, m\} \quad (2.4.5)$$

measuring the error of approximating the true continuous eigenfunctions of the transfer operator,  $\psi_i$  and define

$$\delta := \max_i \delta_i$$

as the largest approximation error amongst these first  $m$  eigenfunctions. The *spectral error*

$$\eta(\tau) := \frac{\lambda_{[m+1]}(\tau)}{\lambda_2(\tau)}$$

is the error due to neglecting the fast subspace of the transfer operator, which decays to zero with increasing lag time:  $\lim_{\tau \rightarrow \infty} \eta(\tau) = 0$ . The general statement is that the Markov model error  $E(k)$  can be bounded [49] from above by the following expression

$$E(k) : = \left\| \mathbb{X} (\mathbf{T}(\tau))^k \mathbb{X} - \mathbb{X} (\mathbf{T}(\tau) \mathbb{X})^k \right\| \quad (2.4.6)$$

$$\leq \min\{2, [m\delta + \eta(\tau)] [a(\delta) + b(\tau)]\} \lambda_2^k \quad (2.4.7)$$

with

$$a(\delta) = \sqrt{m}(k-1)\delta \quad (2.4.8)$$

$$b(\tau) = \frac{\eta(\tau)}{1-\eta(\tau)}(1-\eta(\tau)^{k-1}) \quad (2.4.9)$$

which implies two interesting observations:

1. For long times  $k$ , the overall error decays to zero with  $\lambda_2^k$ , where  $0 < \lambda_2 < 1$ , thus the stationary distribution (recovered as  $k \rightarrow \infty$ ) is always correctly modeled, even if the kinetics are badly approximated. This is a direct consequence of the fact, that the stationary eigenvector, the constant function, was chosen to be part of the projection, which is always true for a partition of unity, and
2. the error during the kinetically interesting time scales consists of a product whose terms contain separately the eigenfunction approximation error and the spectral error. Thus, the overall error can be diminished by choosing a fine discretization (where fine means it needs to well trace the slow eigenfunctions, small  $\delta$ ), and using a large enough lag time  $\tau$ .

Depending on the distribution of eigenvalues, the decay of the spectral error  $\eta(\tau)$  with  $\tau$  might be slow. It is thus interesting to consider a special case of the discretization where  $\delta = 0$ . This is achieved by a Markov model that uses a fuzzy partition with membership functions derived from the first  $m$  eigenfunctions  $\psi_i$  of the transfer operator [75]. From a more practical point of view, this situation can be approached by using a Markov model with  $m_{\text{fine}} \gg m_{\text{macro}}$  states located such that they discretize the first  $m$  eigenfunctions with a vanishing discretization error  $\delta \rightarrow 0$ , and then declaring that we are *only* interested in these  $m$  slowest relaxation processes.

In other words, a Markov model can approximate the kinetics of slow processes *arbitrarily well*, provided the discretization can be made sufficiently fine or improved in a way that continues to minimize the eigenfunction approximation error  $\delta$ . This observation can be rationalized by Eq. (2.2.8) which shows that the dynamics of the transfer operator can be exactly decomposed into a superposition of the stationary and the slow and fast processes.

An important consequence of the  $\delta$ -dependence of the error is that the best partition is not necessarily one which uses a few metastable states. Previous work [53, 51, 65, 52] has focused on the construction of partitions with high metastability (defined as the trace of the transition matrix  $\mathbf{T}(\tau)$ ), e.g. the partition into three states shown in Figure 2.4. i. This approach was based on the idea that the discretized dynamics must be approximately Markovian if the system remained in each partition sufficiently long to approximately lose memory [52]. It can be shown that if a system has  $m$  metastable sets with  $\lambda_m \gg \lambda_{m+1}$ ,

then the most metastable partition into  $m$  sets also minimizes the discretization error [49]. Still, the expression for the discretization error given here has two other profound ramifications: Firstly, even in the case where there exists a strong separation of time scales so the system has clearly  $m$  metastable sets, the discretization error can be reduced *even further* by splitting the metastable partition into more than  $m$  sets which are then not metastable. And secondly, even in the *absence* of a strong separation of time scales, the discretization error can be made arbitrarily small by making the partition finer, especially in transition regions, where the eigenfunctions change most rapidly.

Figure 2.4 illustrates the Markov model discretization error on a two-dimensional three-well example where two slow processes are of interest. The top most panels show a metastable partition into 3 sets. As seen in rows four and five of Figure 2.4, the discretization errors  $\|X^\perp \psi_2\|$  and  $\|X^\perp \psi_3\|$  are large near the transition regions, where the eigenfunctions  $\psi_2(x)$  and  $\psi_3(x)$  change rapidly, leading to a large discretization error. Using a random partition (Figure 2.4, iii) makes the situation worse, but increasing the number of states reduces the discretization error (Figure 2.4, iv), thereby increasing the quality of the Markov model. When states are chosen such as to well approximate the eigenfunctions, a very small error can be obtained with few sets (Figure 2.4, ii)

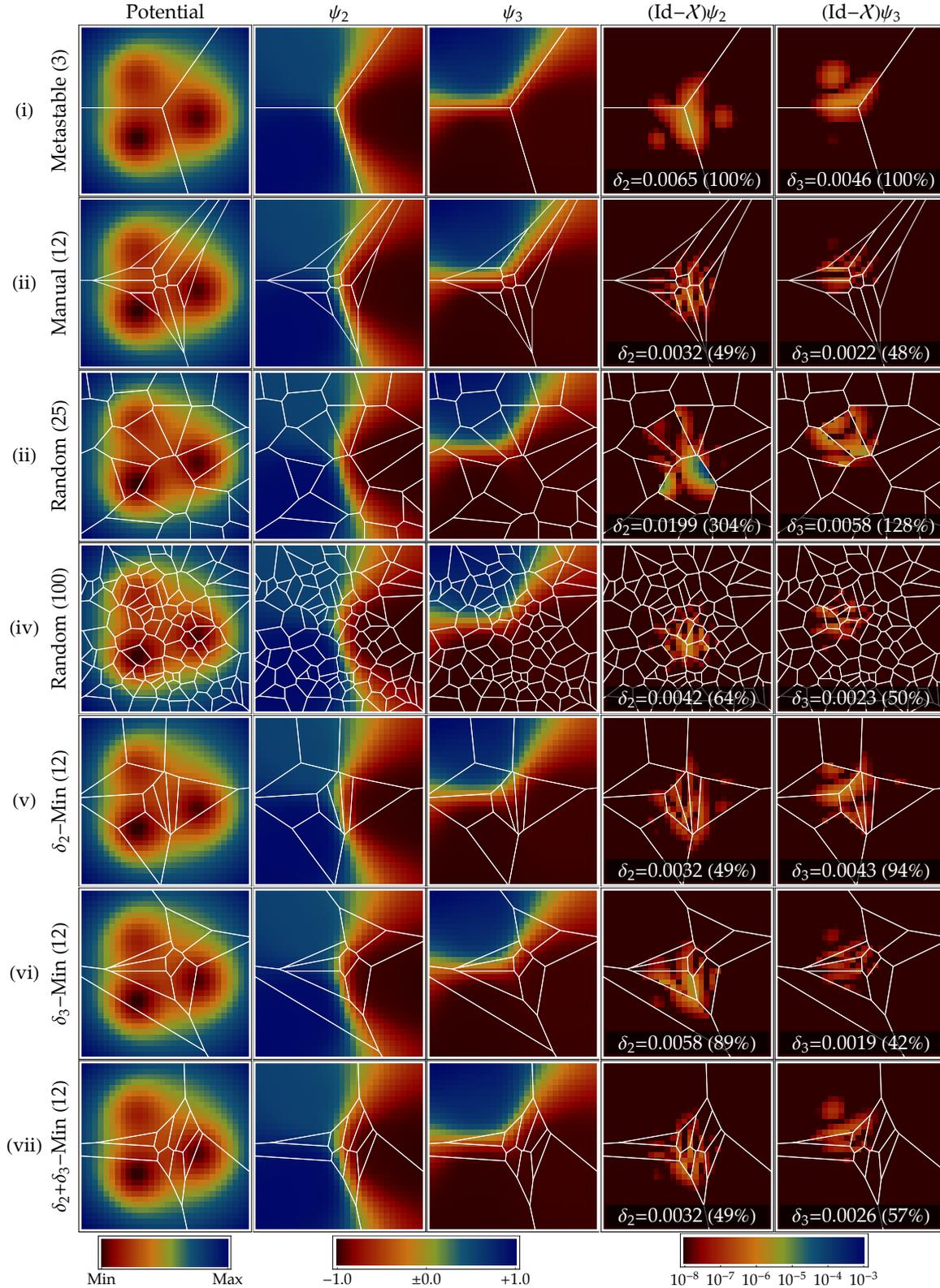
These results suggest that an adaptive discretization algorithm may be constructed which minimizes the  $E(k)$  error. Such an algorithm could iteratively modify the definitions of discretization sets as suggested previously [52], but instead of maximizing metastability it would minimize the  $E(k)$  error which can be evaluated by comparing eigenvector approximations on a coarse discretization compared to a reference evaluated on a finer discretization [49].

For an illustration of this possibility we implemented a simple Metropolis-Monte carlo scheme to optimize the centers  $\mathbf{y}$  of a 12 Voronoi cell partitioning for the previously introduced 2D-model. The initial points were randomly distributed on the entire state space and randomly shifted in each iteration. A metropolis acceptance criterion was chosen to minimize the eigenvector approximation error  $\delta_i$  by

$$\mathbb{P}(\text{accept}) = \min \{1, \exp(-\beta (\delta_i(\mathbf{y}^{\text{new}}) - \delta_i(\mathbf{y}^{\text{old}})))\}$$

where the final minimal solution was chosen after sufficiently long run out of all produced cluster definitions  $\mathbf{y}^{[t]}$ . The result (in Figure 2.4, v-vii) shows that the cluster centers tend to move into the transition region that is important for the selected process.

Combining this idea of discretization with the idea of time scale separation from Eq. (2.2.8) leads to the intriguing insights that if – for a given system – only the slowest dynamical processes are of interest, it is sufficient to discretize the state space in such a way that the first few eigenvectors are well represented (in terms of small approximation errors  $\delta_i$ ). For example, if one is interested in processes on time



**Figure 2.4 – Eigenvector Approximation** Eigenvector approximation errors  $\delta_2$  and  $\delta_3$  on the two slowest processes in a two-dimensional three-well diffusion model (see appendix D for details). Rows: Different state space discretization with white lines as state boundaries: (i) 3 states with maximum metastability, (ii) the metastable states subdivided manually into 12 states with good resolution in the transition region, (iii)/(iv) voronoi partition using 25/100 randomly chosen centers, (v)/(vi)/(vii) optimized centers the position of 12 voronoi cells to minimize  $\delta_2$ ,  $\delta_3$ ,  $\delta_2 + \delta_3$ . Columns: (1) Potential, (2/3) Exact eigenfunctions,  $\psi_2(x)$  and  $\psi_3(x)$ , (4/5) Approximation errors  $X^\perp \psi_2$  and  $X^\perp \psi_3$  with error norms  $\delta_2$  and  $\delta_3$ .

scales  $t^*$  or slower, then the number  $m$  of eigenfunctions that need to be resolved is equal to the number of implied time scales with  $t_i \geq t^*$ . Due to the perfect decoupling of processes for reversible dynamics in the eigenfunctions (see section 2.2), no gap after these first  $m$  time scales of interest is needed, provided that the dynamics of interest occurs within the space spanned by the dominant eigenfunctions. Note, that the quality of the Markov model does not depend on the dimensionality of the simulated system, i.e. the number of atoms. Thus, if only the slowest process of the system is of interest (such as the folding process in a two-state folder), only a one-dimensional parameter, the dominant eigenfunction  $\psi_2(x)$ , needs to be approximated, even if the system is huge. This opens a way to discretize state spaces of very large molecular systems.

### 2.4.3 Approximation of eigenvalues

One of the important and interesting kinetic properties of molecular systems are the *intrinsic time scales*  $t_i$  of the system, the specific time scales it takes the system to relax towards equilibrium. They can be accessed experimentally *via* relaxation or correlation functions that are measurable with various spectroscopic techniques [76, 77, 57, 33].

It is worth noting that observing convergence of the slowest implied time scales in  $\tau$  is not a test of Markovianity. While Markovian dynamics implies constancy of implied time scales in  $\tau$  [51, 53], the reverse is not true and would require the eigenvectors to be constant as well. However, observing the lag time-dependence of the implied time scales is a useful approach to choose a lag time  $\tau$  at which a projected Markov model  $T^X_\tau$  shall be calculated, but this model needs to be validated subsequently (see section 2.6.6).

For Markov models these intrinsic time scales are identified with the eigenvalues, which need to be corrected for the lag time  $\tau$  considered

$$t_i(\tau) = -\frac{\tau}{\log \lambda_i(\tau)}$$

and are constant if the Chapman-Kolmogorov property holds. In this case the relation

$$\lambda_i(\tau) = \bar{\lambda}_i^\tau$$

is true with lag time independent eigenvalues  $\bar{\lambda}_i \equiv \lambda_i(\tau_0)$  chosen at some native time-step  $\tau_0$  and conclusively

$$\begin{aligned} t_i(\tau) &= -\frac{\tau}{\log \lambda_i(\tau)} \\ &= -\frac{\tau}{\log \bar{\lambda}_i^\tau} \\ &= -\frac{1}{\log \bar{\lambda}_i} = \text{const.} \end{aligned}$$

If we now construct a Markov model from the observed process which is not Markovian the ITS can be used as a measure of how well the Markov property holds. More precisely, we can only infer a good Markov model, that is a model for some projected Markovian dynamics, if the ITS  $t_i$  is constant not vice versa. Thus, it is desirable to find a measure, how well the eigenvalues of the projected model will approximate the exact eigenvalues. Again, we consider the first  $m$  eigenvalues of the transfer operator  $T_\tau$ ,  $1 = \lambda_1(\tau) > \lambda_2(\tau) \geq \dots \geq \lambda_m(\tau)$ , and let  $1 = \hat{\lambda}_1(\tau) > \hat{\lambda}_2(\tau) \geq \dots \geq \hat{\lambda}_m(\tau)$  denote the associated eigenvalues of the Markov model  $T^X$ . The rigorous mathematical estimate from [78] then states that

$$\max_{j=2,\dots,m} |\lambda_j(\tau) - \hat{\lambda}_j(\tau)| \leq (m-1) \lambda_2(\tau) \delta^2, \quad (2.4.10)$$

where  $\delta$  is the previously introduced maximum discretization error of the first  $m$  eigenfunctions. This shows that the eigenvalues are well reproduced when the discretization traces these eigenfunctions well. In particular if we are only interested in the eigenvalue of the slowest process,  $\lambda_2(\tau)$ , which is often experimentally reported *via* the slowest relaxation time of the system,  $t_2$ , the following estimate of the approximation error can be given:

$$\frac{|\lambda_2(\tau) - \hat{\lambda}_2(\tau)|}{|\lambda_2(\tau)|} \leq \delta_2^2 \quad (2.4.11)$$

As  $\lambda_2(\tau)$  corresponds to a slow process, we can make the restriction  $\lambda_2(\tau) > 0$ . Moreover, the discretization error of Markov models based on full partitions of state space is such that the eigenvalues are always underestimated [78], thus  $\lambda_2(\tau) - \hat{\lambda}_2(\tau) > 0$ . Using Eq. (2.2.7), we obtain the estimate for the discretization error of the largest implied time scale:

$$\hat{t}_2^{-1} - t_2^{-1} \leq -\tau^{-1} \ln(1 - \delta_2^2), \quad (2.4.12)$$

which implies that for either  $\delta_2 \rightarrow 0_+$  or  $\tau \rightarrow \infty$ , the error in the largest implied time scale tends to zero. Moreover, since  $\lambda_2(\tau) \rightarrow 0$  for  $\tau \rightarrow \infty$ , this is also true for the other processes:

$$\lim_{\tau \rightarrow \infty} \frac{|\lambda_j(\tau) - \hat{\lambda}_j(\tau)|}{|\lambda_j(\tau)|} = 0, \quad (2.4.13)$$

and also

$$\lim_{\delta \rightarrow 0} \frac{|\lambda_j(\tau) - \hat{\lambda}_j(\tau)|}{|\lambda_j(\tau)|} = 0. \quad (2.4.14)$$

which means that the error of the implied time scales also vanishes for either sufficiently long lag times  $\tau$  or for a sufficiently fine discretization. This fact has been empirically observed in many previous studies [79, 53, 52, 58, 44, 51, 35], but can now be understood in detail in terms of the discretization error.

The implications of these two theorems are important: We now have the mathematical justification to express the dynamics of a system with continuous state space by an approximation or stated in reverse for each system it is possible to find a discrete approximation for any given initial level of accuracy.

## 2.5 DISCRETE STATE SPACE / TRANSITION MATRIX

After a thorough discussion of the problems involved in projections, we are now making the transition to the case of a discrete state space or the projection of a (possibly very large) discretized space. With the results above it is – at least for the physical relevant cases – possible to find a discretization, that is sufficiently fine to reproduce the qualities of the original Markov model arbitrarily well.

Since the propagation of the projected operator  $T^X$  is closed with respect to the projection  $X$  we can express the propagation in terms of the basis  $\chi$  that spans this closed subspace of  $\Omega$ . Provided that the number of macro states,  $|\mathbf{M}| = M < \infty$ , is finite, we will define the *transition matrix*  $\mathbf{T}(\tau) \in \mathbb{R}^{M \times M}$  by

$$T_{ij}(\tau) = \frac{\langle \chi_j, T_\tau \chi_i \rangle}{\langle \chi_i, \chi_i \rangle}.$$

In the case of a crisp partitioning, using the definition for the transfer operator in Eqs. (2.2.1) and (2.2.2), this is equal to

$$\begin{aligned} T_{ij}(\tau) &= \frac{\int_{\Omega} \mathbf{d}x \pi(x) \chi_j(x) [\Pi^{-1} \mathbf{Q}_\tau \Pi \chi_i](x)}{\int_{\Omega} \mathbf{d}x \pi(x) \chi_i(x) \chi_i(x)} \\ &= \frac{\int_{\Omega} \mathbf{d}x \pi(x) \chi_j(x) \pi^{-1}(x) \int_{\Omega} \mathbf{d}y p_\tau(x, y) \pi(y) \chi_i(y)}{\int_{\Omega} \mathbf{d}x \pi(x) \chi_i(x) \chi_i(x)} \\ &= \frac{\int_{\Omega} \mathbf{d}x \chi_j(x) \int_{\Omega} \mathbf{d}y p_\tau(x, y) \pi(y) \chi_i(y)}{\int_{\Omega} \mathbf{d}x \pi(x) \chi_i(x) \chi_i(x)}. \end{aligned}$$

Choosing the indicator functions  $\mathbf{1}_{\omega_i}(x)$  for the crisp clustering  $\omega_i$  as membership functions  $\chi_i(x)$

$$\chi_i(x) = \mathbf{1}_{\omega_i}(x) \equiv \begin{cases} 1 & \text{if } x \in \omega_i \\ 0 & \text{else} \end{cases}$$

the matrix simplifies to

$$\begin{aligned} T_{ij}(\tau) &= \frac{\int_{\omega_j} \mathbf{d}x \int_{\omega_i} \mathbf{d}y p_\tau(x, y) \pi(y)}{\int_{\omega_i} \mathbf{d}x \pi(x)} \\ &= \frac{\mathbb{P}[x_{t+\tau} \in \omega_j \wedge x_t \in \omega_i]}{\mathbb{P}[x_t \in \omega_i]} \\ &= \mathbb{P}[x_{t+\tau} \in \omega_j \mid x_t \in \omega_i] \end{aligned}$$

which corresponds to the counting of relative transition probabilities between subsets  $\omega_i$  and  $\omega_j$  in an infinitely long trajectory and is thus exactly what we desired it to be. Note that in this crisp case the integrals run over individual sets and only need the local equilibrium distributions

$$\pi_i \equiv \int_{\omega_i} dx \pi(x)$$

as weights. This is a very powerful feature: In order to estimate transition probabilities, we do not need any information about the global equilibrium distribution of the system, and the dynamical information needed extends only over time  $\tau$ . In principle, the full dynamical information of the discretized system can be obtained by initiating trajectories of length  $\tau$  out of each state  $i$  as long as we draw the starting points of these simulations from a local equilibrium density [75, 56, 55] given by

$$\begin{aligned} \pi^{[\omega_i]}(x) &= \delta(x \in \omega_i) \frac{\pi(x)}{\langle \delta(x \in \omega_i), \delta(x \in \omega_i) \rangle} \\ &= \delta(x \in \omega_i) \frac{\pi(x)}{\pi_i}. \end{aligned}$$

Lastly, the transition matrix can also be written in terms of *correlation functions*

$$\text{cor}[a | b](\tau) \equiv \mathbb{E}_t [a(x_t) b(x_{t+\tau})] \quad (2.5.1)$$

of the time series  $x_t$  [53] between two state-space observables  $a, b : \Omega \mapsto \mathbb{R}$ . Using two membership functions  $\chi_i$  and  $\chi_j$  from a (fuzzy or crisp) clustering  $\chi$  will then measure the absolute probability to find a transition between the two substates  $i$  and  $j$  in the observed trajectory  $x_t$ . Correctly normalized these resemble the conditional jump probabilities of the transition matrix  $\mathbf{T}$  that are then given by

$$T_{ij}(\tau) = \frac{\text{cor}[\chi_i | \chi_j](\tau)}{\mathbb{E}_t[\chi_i]}. \quad (2.5.2)$$

and for a crisp clustering even

$$T_{ij}(\tau) = \frac{\text{cor}[\chi_i | \chi_j](\tau)}{\text{cor}[\chi_i | \chi_i](\tau)}. \quad (2.5.3)$$

The use of correlation functions will be addressed in more detail in chapter 5.

From now on we will assume w.l.o.g. that the exact underlying Markovian dynamics can be approximated well enough by a finite state space. Since the transition matrix  $\mathbf{T}(\tau)$  is a discretization of the transfer operator  $T_\tau$  [72, 49, 55] (see section 2.2) and we explicitly assume detailed balance (see section 2.1.1), we can relate the functions  $v_t$  and probability densities  $\mathbf{p}_t$  that are transported by  $T_\tau$  and  $Q_\tau$  to

Chapman-Kolmogorov semi-group property
$\mathbf{T}(k \cdot \tau) = [\mathbf{T}(\tau)]^k$
Eigensystem decomposition of transition matrix
$\mathbf{T}(\tau) = \mathbf{R}\mathbf{\Lambda}(\tau)\mathbf{R}^{-1} = \mathbf{L}^{-\mathbf{T}}\mathbf{\Lambda}(\tau)\mathbf{L}^{\mathbf{T}}$
normalization convention
$\mathbf{R}\mathbf{L} = \text{Id}$
with detailed balance $\mathbf{\Pi}\mathbf{T} = \mathbf{T}^{\mathbf{T}}\mathbf{\Pi}$
$\mathbf{L} = \mathbf{R}^{\mathbf{T}}\mathbf{\Pi}$
$\mathbf{L}^{\mathbf{T}}\mathbf{L} = \mathbf{\Pi}, \mathbf{R}\mathbf{R}^{\mathbf{T}} = \mathbf{\Pi}^{-1}$
$\mathbf{T}(\tau) = \mathbf{R}\mathbf{\Lambda}(\tau)\mathbf{L} = \mathbf{R}\mathbf{\Lambda}(\tau)\mathbf{R}^{\mathbf{T}}\mathbf{\Pi} = \mathbf{\Pi}^{-1}\mathbf{L}^{\mathbf{T}}\mathbf{\Lambda}(\tau)\mathbf{L}$
Correlation matrices
$\mathbf{C}(\tau) = \text{cor}[\mathbf{X}_i, \mathbf{X}_j](\tau)$
for the identity clustering $\mathbf{X} = \text{Id}$
$\mathbf{C}(\tau) = \mathbf{\Pi}\mathbf{T}(\tau)$

**Table 2.1 – Matrix Relations** used for discrete time, discrete state space Markov State Model (MSM)

column and row vectors that are multiplied to the transition matrix. The choice of  $\mathbf{T}(\tau)$  being row-stochastic corresponds to

$$\mathbf{p}_t \mapsto \mathbf{p}^{\mathbf{T}}(t)$$

as row vectors that are multiplied to the matrix from the left and

$$\mathbf{v}_t \mapsto \mathbf{v}(t)$$

as column vectors from the right [79, 53, 51, 80, 81, 44, 52, 73, 49]. This finally allows to simplify the notation to matrix operations in the following chapters. We will use the symbols given in Table A.1 in the remainder of this thesis and assume the relations in Table 2.1, which are a direct consequence of the concepts introduced for the transfer operator and the propagator.

### 2.5.1 Discretization Methods for molecules

Macromolecular systems generally possess configuration spaces of such high dimension that grid-based methods for partitioning space become impractical, why this is often referred to as the “Curse of Dimensionality”. However, in many macromolecular systems such as proteins, the region over which the configurational probability density is significant defines a low-dimensional (but potentially highly nonlinear) subspace [82]. As a result, data-driven methods, where a clustering of conformations sampled by some form of molecular simulation defines the partitioning of this low-dimensional subspace, are

both attractive and practical. Various combinations of distance metrics and clustering methods have been proposed: Distance metrics include Euclidean distance in backbone coordinates [35] or Root Mean Squared Displacement (RMSD) [62, 52]. Clustering methods include manual clustering [83], k-means clustering [35], k-centers clustering [62], density-based clustering [84, 85] and adaptive clustering approaches [52]. Approaches to directly discretize certain coordinates, such as the rotameric states [64, 44] or the hydrogen-bond patterns [51, 79] were also made.

All of these metrics and clustering methods have their advantages and disadvantages and the applicability and feasibility has to be decided from case to case, there is no best or worst method. Every method that allows a more accurate approximation of the (dominant) eigenvectors with increasing number of clusters permits the generation of a Markov model to arbitrary precision. A metric or a clustering algorithm that optimizes the macro states w.r.t. this criterion is difficult to design and so, in practice, one often measures structural differences on a subset of coordinates (e.g. backbone coordinates). In almost any practical way, the approximation is not perfect and the eigenvector approximation error will not decrease to zero. That must be compensated by increasing the lag time  $\tau$ . Generally, it is important that the metric is selected such that the molecular events under investigation can be resolved.

However, it is interesting to see that MSMs are robust with respect to changes of the metric and the clustering method, within a significant range[1].

Markov models typically only use atom positions while the velocities are projected out [52, 51]. So far, Markov models have also neglected solvent degrees of freedom and have only used the solute coordinates [52, 35], and the effect of this was studied in detail in [84]. Indeed, it may be necessary to incorporate solvent coordinates in situations where the solvent molecules are involved in slow processes that are not easily detected in the solute coordinates [86]. Often, Markov models are also based on distance metrics that only involve a subset of the solute atoms, such as RMSD between heavy atom or alpha carbon coordinates [52, 62, 35], or backbone dihedral angles [51, 44]. Possibly the strongest approximation is caused by clustering or lumping sets of coordinates in the selected coordinate subspace into discrete states [79, 52, 44, 62, 35].

## 2.6 ESTIMATION FROM DATA AND VALIDATION

So far the continuous transfer operator has been the basis for the continuous and discretized Markov models. In almost all practical cases we have no access to the transfer operator or any other analytical de-

scription and the transition matrix  $\mathbf{T}(\tau)$  has to be estimated from a finite quantity of simulation or even experimental data.

This includes a *statistical error* component into the overall error in modeling the true dynamics with Markov models which will be discussed in this section. Here we assume that a sufficiently accurate state space discretization (either crisp or fuzzy) has been defined and that a trajectory is projected onto this discrete space. We then address the question how to estimate a Markov model based on such trajectory data. Note that while in the previous section we have studied only the discretization error of the Markov model without consideration of statistical issues (i.e., it was assumed that the transition matrix could be computed exactly), this section *only* studies statistical issues without consideration of the discretization error (i.e. the discrete dynamics is now assumed to be perfectly Markovian).

### 2.6.1 From trajectories to count matrix

We start with a time series (a trajectory) which is observed at fixed time intervals  $\Delta t$  at  $L + 1$  points

$$\mathbf{t} : t_k = k\Delta t, \quad k \in \{0, \dots, L\}$$

and use the simplified discrete trajectory

$$x_k : k \in \{0, \dots, L\} \mapsto x(t_k) \in \Omega$$

as before for a finite state space  $\Omega = \{1, \dots, M\}$ . For convenience, integer variables are used for the time with equidistant time intervals. The state-to-state correlations in Eq. 2.5.2 have been used to compute the transition probabilities since these can measure the absolute probabilities for a transition in a time series. To get an estimate of the state-to-state correlation from a given time series  $x_k$  we can use the definition in Eq. 2.5.1 and get

$$\begin{aligned} \text{cor} [\chi_i | \chi_j] (l) &\equiv \mathbb{E}_n [\chi_i(x_n) \chi_j(x_{n+l})] \\ &= (L + 1 - l)^{-1} \sum_{n=0}^{L-l} \chi_i(x_n) \chi_j(x_{n+l}) \end{aligned}$$

for finite time series. This naturally leads to the *count matrix*  $\mathbf{Z}$  defined to consist of the absolute number of transitions in the trajectory when using a fixed lag time  $l$  by

$$Z_{ij}(l) = \sum_{n=0}^{L-l} \chi_i(x_n) \chi_j(x_{n+l})$$

or simply

$$Z_{ij}(l) = (L + 1 - l) \text{cor} [\chi_i | \chi_j] (l\Delta t)$$

leaving a total of  $L$  transition counts for the shortest possible lag time  $l = 1$ . As a shorthand notation we define the row sums of  $\mathbf{Z}$  by

$$z_i(l) \equiv \sum_{j \in \mathcal{M}} Z_{ij}(\tau), \quad (2.6.1)$$

which are the total number of times the trajectory started from state  $i$ . Note, that  $\mathbf{z}(l)$  is not independent of the lag time  $l$  since the last  $l$  timesteps at the end of the trajectory are omitted from the counting.

The purpose of this counting is two-fold: To estimate transition probabilities and their statistical uncertainty in these estimates. While the estimation (so far) does not depend on the total number of observed transitions, the statistical properties, such as standard deviations surely will. Regarding this, one has to be careful when *counting*.

We distinguish between two approaches to counting, i.e to estimate the number of events of interest from an observed trajectory. Both approaches differ only in the case where we estimate for lag times  $l > 1$  larger than the native sampled time step  $\Delta t$ . The remaining question is, how to treat *jumps between jumps*. Nevertheless, the total number of counts has to match the total length of the trajectory  $t^{\text{total}} = L\Delta t$  divided by the lag time  $\tau = l\Delta t$  which is the number of jumps of lag time  $\tau$  that can be contained in the trajectory.

#### A. Sampling at fixed time intervals $\tau$ only

Here the trajectory is sampled at lag time intervals  $\tau$  and only these sample points are used for counting

$$Z_{ij}(l) = \sum_{k=0}^{\lfloor L/l \rfloor - 1} \chi_i(x_{l \cdot k}) \chi_j(x_{l \cdot (k+1)}) \quad (2.6.2)$$

while intermediate points are neglected. When the observed jump process is Markovian at  $\tau$ , this generates statistically independent transition counts. It is therefore straightforward to use the resulting count matrix in order to derive expressions for the likelihood and posterior of the transition matrix (see sections below). This is important in order to obtain models that do not underestimate the statistical uncertainties [58, 52, 64]. A disadvantage of this approach is that for lag times  $l \gg 1$  much of the data is ignored, which can lead to numerical problems. In particular, states that have been actually visited or transitions that have been actually observed might be missed when subsampling the data at interval  $\tau$ , which may be a reason for estimators breaking down or giving problems to compare estimations at different lag times.

#### B. Sampling using a shifting window

In this method we use a count window of width  $\tau = l\Delta t$  that is shifted along the time line

$$Z_{ij}(l) = \frac{(L+1-l)}{l} \text{Cor} [\chi_i, \chi_j] (l\Delta t)$$

which is exactly the way the computation of transitions using a correlation function works with the benefit that this method uses all observed transitions. The resulting count matrix has to be corrected for *over-counting* with a factor of  $1/l$  as for lag times  $l > 1$  not all counts can be considered independent. However, the scaling does not affect maximum posterior estimators (section 2.6.3), such that the estimates are asymptotically correct and the window count method is generally preferred for this case. In principle this counting can be regarded as an averaging over many estimations with the same length: The error stays, but the mean improves. If the observations would be independent, which we can normally not assume, also the error would decrease!

### 2.6.2 Likelihood, Bayesian

We finally address the problem of statistical uncertainty. The objective is to find a parametrization of a Markov State Model that is able to reproduce the dynamics in the trajectory. It is intuitively clear that in the limit of an infinitely long trajectory, the elements of the true transition matrix are given by the trivial estimator

$$\hat{T}_{ij}(\tau) = \frac{Z_{ij}(\tau)}{\sum_{k \in \mathcal{M}} Z_{ik}(\tau)},$$

i.e. the fraction of times the transition  $i \rightarrow j$  led out of state  $i$  into state  $j$ . For a trajectory of limited length, the transition matrix  $\mathbf{T}$  that reproduces this trajectory is no longer uniquely determined which can easily be seen: For a trajectory of length  $L+1$  the above estimator would still result in a valid transition matrix, with the additional property that it will reproduce the initial observation more likely than any other transition matrix. Still, small (or even large) changes in these transition probabilities might also, although being less probably, lead to the same observation and so the estimation from a finite trajectory must carry an uncertainty.

The better the agreement, the more the predictions from the model can be assumed to be correct. Here, we introduce a term called *likelihood*  $\mathcal{L}$  which is the probability, that a given observation (e.g. a trajectory) will be reproduced by a certain (Markov) model, in our case, represented by the sought-after transition matrix  $\mathbf{T}$ . Assuming, that a parametrized model is already given by a transition matrix  $\mathbf{T}$ , we can, following [87], write down the *likelihood*  $\mathbb{P}(\mathbf{Z} | \mathbf{T})$  for this particular model to reproduce a certain observation that is encoded in a particular count matrix  $\mathbf{Z}$ . As long as the matrix  $\mathbf{Z}$  contains statistically independent transition counts (see discussion in section 2.6.1

above), the probability that a particular  $\mathbf{T}$  will generate a sequence  $\{y_0, \dots, y_L\}$  as the observed trajectory is given by the product

$$\mathbb{P}(\mathbf{Z} | \mathbf{T}) \propto \prod_{k=0}^{L-1} T_{[y_k, y_{k+1}]}$$

of the individual jump probabilities  $T_{ij}$ . The result in terms of the count matrix  $\mathbf{Z}$  is a multinomial distribution

$$\mathbb{P}(\mathbf{Z} | \mathbf{T}) \propto \prod_{i,j \in \mathcal{M}} T_{ij}^{Z_{ij}} \quad (2.6.3)$$

or in the more convenient logarithmic likelihood

$$\log \mathbb{P}(\mathbf{Z} | \mathbf{T}) = \sum_{i,j \in \mathcal{M}} Z_{ij} \log(T_{ij}) + \text{const}$$

where we omitted the normalization parameters. What we were actually aiming for is the reverse direction of implication, the posterior probability  $\mathbb{P}(\mathbf{T} | \mathbf{Z})$ : The answer to the question, How probable is a particular model, if the observation is given? To solve this, Bayes' theorem (found e.g. in [88]) is applied which states that

$$\begin{aligned} \mathbb{P}(\mathbf{T} | \mathbf{Z}) &= \frac{\mathbb{P}(\mathbf{Z} | \mathbf{T})\mathbb{P}(\mathbf{T})}{\mathbb{P}(\mathbf{Z})} \\ &= \frac{\mathbb{P}(\mathbf{Z} | \mathbf{T})\mathbb{P}(\mathbf{T})}{\sum_{\mathbf{T}'} \mathbb{P}(\mathbf{Z} | \mathbf{T}')\mathbb{P}(\mathbf{T}')} \end{aligned}$$

or even simpler

$$\mathbb{P}(\mathbf{T} | \mathbf{Z}) \propto \mathbb{P}(\mathbf{Z} | \mathbf{T})\mathbb{P}(\mathbf{T}).$$

All we have to do is to choose  $\mathbb{P}(\mathbf{T})$ , the *prior* probability to consider a certain model in the first place without any knowledge about an observation. We explicitly assume that the parametrization of  $\mathbf{T}$  only permits valid transition matrices and that we do not need to use the prior probabilities for this. The simplest choice of a prior would then be to assume all models to be equally likely

$$\mathbb{P}^{\text{uniform}}(\mathbf{T}) \propto 1$$

which seems trivial, but leads to Dirichlet distributions as the wanted posterior probability distribution  $\mathbb{P}(\mathbf{T} | \mathbf{Z})$ . If we want to include more a priori knowledge about our model we can do this, although it is often practical to use a *conjugate prior* to the multinomial likelihood distribution. This ensures that prior and posterior will have the same (algebraic) form which is a Dirichlet Distribution for the multinomial model parametrization in the likelihood  $\mathcal{L}$ .

If the prior is given by a Dirichlet distribution we can write down the form

$$\log \mathbb{P}(\mathbf{T}) = \sum_{i,j \in \mathcal{M}} Z_{ij}^{\text{prior}} \log(T_{ij}) + \text{const}$$

which can be regarded (just from the algebraic shape) as a likelihood function for a transition matrix with the same transition probabilities but *additional counts*  $\mathbf{Z}^{\text{prior}}$ . Now using different priors is equivalent to simply adding *prior counts* to the real observed ones

$$\log \mathbb{P}(\mathbf{T} | \mathbf{Z}) = \sum_{i,j \in \mathcal{M}} \left( Z_{ij}^{\text{prior}} + Z_{ij} \right) \log(T_{ij}).$$

As long as we use a prior of this form, we define the effective count matrix by

$$\mathbf{Z}^{\text{eff}} := \mathbf{Z}^{\text{prior}} + \mathbf{Z}.$$

and note that the case of the uniform prior  $\mathbb{P}(\mathbf{T}) = 1$  implies that there are no prior counts  $Z_{ij}^{\text{prior}} = 0$  which is in agreement with the interpretation as *additional counts*.

### 2.6.3 Maximum probability estimators

Once we have a computable form of the posterior, we can find representative models with certain properties from the total set of possible parameters. The application of some of these properties leads to an algebraic form of the transition matrix in terms of the count matrix. If we assume a uniform prior and want the model with the highest posterior probability, we find

$$\hat{T}_{ij}(\tau) = \left[ \underset{\mathbf{T}}{\operatorname{argmax}} \mathbb{P}(\mathbf{T} | \mathbf{Z}) \right]_{ij} = \frac{Z_{ij}(\tau)}{\sum_{k \in \mathcal{M}} Z_{ik}(\tau)} \quad (2.6.4)$$

for the estimated transition probabilities, which is simply the row-wise normalized count matrix or the estimation we have computed from the canonical correlation matrix.

In the limit of infinite sampling, i.e., trajectory length  $N \rightarrow \infty$ ,  $\mathbb{P}(\mathbf{T} | \mathbf{C})$  converges towards a Dirac delta distribution with its peak at  $\hat{\mathbf{T}}(\tau)$ . In this case the prior contribution vanishes

$$\lim_{N \rightarrow \infty} \hat{T}_{ij} = \lim_{N \rightarrow \infty} \frac{Z_{ij}^{\text{prior}} + Z_{ij}}{Z_i^{\text{prior}} + z_i} = \lim_{N \rightarrow \infty} \frac{Z_{ij}}{z_i} = T_{ij}, \quad (2.6.5)$$

i.e., the estimator is “*asymptotically unbiased*”.

Note, that the estimator  $\hat{\mathbf{T}}(\tau)$  does not necessarily fulfill detailed balance  $\mathbf{\Pi} \hat{\mathbf{T}} = \hat{\mathbf{T}}^T \mathbf{\Pi}$  even if the underlying dynamics is in equilibrium and thus detailed balance holds for the true transfer operator  $T_\tau$ . In many cases it is desirable and advantageous to estimate a transition matrix that does fulfill detailed balance, but there is no known closed form solution for the maximum probability estimator with the detailed balance constraint. One simple way to ensure detailed balance for trajectories sampled from equilibrium is to count the trajectory

forward and backwards and correct for the over-counting. The count matrix then takes a symmetric form

$$\mathbf{Z}^{\text{sym}}(\tau) = \frac{1}{2}(\mathbf{Z} + \mathbf{Z}^T)$$

and as long as the prior is symmetric, the posterior estimator will also fulfill detailed balance [53, 79]. This method is asymptotically correct if the underlying dynamics is reversible and the starting points of the trajectory were drawn from equilibrium. In this case forward and backward counted trajectory have equal probabilities and so the combined use leads to a symmetric count matrix sampled from the same dynamics. Another, more direct way was presented in Ref. [62], an iterative method to obtain a reversible estimator that was later reformulated into a computationally more efficient algorithm [1].

#### 2.6.4 Expectation and Uncertainty

Since simulation data is finite, all validation procedures (either consistency checks or comparisons to experimental data) need to account for statistical uncertainties. For these, standard deviations or confidence intervals induced by the posterior distribution of transition matrices are of interest. It follows from the well-studied properties of the Dirichlet distribution in the posterior distribution of transition matrices [87] that the expectation value for posterior distribution of transition matrices is given by

$$\bar{T}_{ij} = \mathbb{E}[T_{ij}] = \frac{Z_{ij}^{\text{eff}} + 1}{z_i^{\text{eff}} + M}$$

while the variance takes the form

$$\text{Var}[T_{ij}] = \frac{(Z_{ij}^{\text{eff}} + 1)((z_i^{\text{eff}} + M) - (Z_{ij}^{\text{eff}} + 1))}{(z_i^{\text{eff}} + M)^2((z_i^{\text{eff}} + M) + 1)} = \frac{\bar{T}_{ij}(1 - \bar{T}_{ij})}{z_i^{\text{eff}} + M + 1}.$$

Both equations display a  $1/M$  dependency. To understand this implication consider a trajectory of a given molecular system which is analyzed with two different state space discretization, one with  $M^{\text{small}} = 10$  and one with  $M^{\text{large}} = 1000$  and assume that one lag time  $\tau$  has been chosen which is long enough to provide Markov models with small discretization error for both  $M$  (discussed in the previous section). When using a uniform prior ( $\mathbf{Z}^{\text{eff}} = \mathbf{Z}$ ), the expectation values would be different for the two discretization: In the  $M^{\text{large}}$  case, if the total number of observations  $L$  is not large enough, the information is small compared to the information inferred by the prior distribution and the expectation value would be biased towards the uninformative  $T_{ij} \approx 1/M$  matrix. Compared to the case of  $M^{\text{small}}$  case one would need  $(M^{\text{large}}/M^{\text{small}})^2$  times more samples to similarly suppress the prior. This behavior is statistically correct w.r.t. to

the prior assumption, but for practical cases undesirable. Thus, for uncertainty estimations it is suggested to use a less influencing prior which allows the observed data to have more impact. The general proceeding is to keep the prior uniform to not introduce an artificial bias  $Z_{ij}^{\text{prior}} = \alpha$  and reduce the value of the *contraction parameter*  $\alpha \rightarrow -1$ . The result is, that the prior probability distribution (without observations) is shifted from a totally uniform distribution ( $\alpha = 0$ )

$$\mathbb{P}(\mathbf{T}^{\text{prior}}) = \text{const}$$

towards a highly localized distribution ( $\alpha \gtrsim -1$ ) where only one entry per row is approximately one and the others are about zero

$$\mathbb{P}(\mathbf{T}^{\text{prior}}) \propto \begin{cases} 1 & \text{if } \forall a \exists b : T_{ab} \approx 1, \quad T_{[a,c \neq b]} \approx 0 \\ 0 & \text{else} \end{cases}$$

Three choices are common:

1. the  $1/M$  prior ( $\alpha = 1/M - 1$ ) leading to the first moments [[2, 60]]

$$\bar{T}_{ij} = \mathbb{E}[T_{ij}] = \frac{Z_{ij} + 1/M}{z_i + 1},$$

$$\text{Var}[T_{ij}] = \frac{(Z_{ij} + 1/M)((z_i + 1) - (Z_{ij} + 1/M))}{(z_i + 1)^2((z_i + 1) + 1)} = \frac{\bar{T}_{ij}(1 - \bar{T}_{ij})}{z_i + 2},$$

2. the “Null prior” ( $\alpha = -1$ ), the extreme case [35, 2] (see chapter 3) giving the moments

$$\bar{T}_{ij} = \mathbb{E}[T_{ij}] = \frac{Z_{ij}}{z_i} = \hat{T}_{ij},$$

$$\text{Var}[T_{ij}] = \frac{(Z_{ij})(z_i - Z_{ij})}{(z_i)^2(z_i + 1)} = \frac{\bar{T}_{ij}(1 - \bar{T}_{ij})}{z_i + 1}.$$

Thus, with a null prior, the expectation value is located at the likelihood maximum, or equivalently at the maximum of the posterior when a uniform prior would be used.

3. The “neighborhood prior” which assumes additional knowledge about the topological structure of the model and assumes a symmetric (e.g. uniform or  $1/M$  prior) restricted to the subset of transitions induced by the topology (here the states assumed to be adjacent within the lag time  $\tau$ ) and a “null prior” on all other transitions. An example will be given in chapter 3.

For the first two prior choices, the variances  $\text{Var}[T_{ij}]$  are independent of the number of discretization bins used and all three of them decay asymptotically with the number of transitions  $Z_{i\text{out}}$  of the state  $i$ , which is expected for sampling expectations from the central limit theorem.

### 2.6.5 Errors in Dependent Variables

In practice, one is often not primarily interested in the uncertainties of the transition matrix elements themselves, but rather in the uncertainties in properties  $f(\mathbf{T})$  computed *from* the transition matrix. Two different approaches have been suggested for this:

1. **Linear error perturbation** [60, 2, 61]: Here, the posterior distributions of transition matrices are approximated by a Multivariate Gaussian Distribution using the first two moments. If the property of interest  $f(\mathbf{T})$  is approximated by a first-order Taylor expansion

$$f(\mathbf{T}) \approx f(\bar{\mathbf{T}}) + \sum_{i,j \in \mathbf{M}} (T_{ij} - \bar{T}_{ij}) \partial_{T_{ij}} f \Big|_{\mathbf{T}}$$

the distribution in transition matrices can be propagated using Gaussian error propagation. Here,  $\partial_{T_{ij}} f \Big|_{\mathbf{T}}$  refers to the partial derivative of the function  $f$  w.r.t. the  $i, j$ -th entry in the transition matrix  $\mathbf{T}$  evaluated at the mean transition matrix  $\bar{\mathbf{T}}$ . This results in a Gaussian distribution of the property of interest with a mean and a covariance matrix that can be computed in terms of  $\mathbf{Z}$ . The approach is deterministic, which is desirable in situations where uncertainties are used to steer an adaptive sampling procedure [89, 50, 61, 60], and may be implemented very efficiently. The disadvantage of this approach is that the Gaussian approximation of the transition matrix posterior is only asymptotically valid, but breaks down easily when few counts have been observed and hence permits unphysical values (e.g.  $T_{ij}$  outside the range  $[0,1]$ ). Moreover, depending on the non-linearity of  $f$ , the first-order Taylor expansion is only valid for small deviations from the mean and thus the error estimation is only good for large numbers of transitions at all.

2. **Markov chain Monte Carlo (MCMC) sampling of transition matrices** [90, 64, 58]: In this approach, a set of transition matrices is drawn from the posterior distribution. The property of interest is then calculated for each transition matrix, and the uncertainties are directly estimated from this set. This approach requires that the true distribution is sampled often enough so that well-converged estimates of standard deviations or confidence intervals can be made. The advantage of this approach is that no approximations are made concerning the functional form of the distribution or the property being computed. Furthermore, this approach can be straightforwardly applied to any function or property of transition matrices, including complex properties such as transition path distributions [35] without deriving the expressions necessary for the linear error perturbation analysis.

Its disadvantage is that sampling may become slow for large matrices or complex objective functions.

The issue will be addressed in detail in chapter 3.

### 2.6.6 Validation: Chapman-Kolmogorov test

As mentioned earlier it is often usually not possible to describe the dynamics in a projected / discretized observable by a Markov model exactly. And as the choice of an appropriate lag time for the parametrization of our model is not unique it is desirable to find a suitable measure for these purposes. The goal is to match the prediction of the dynamics as closely to the observation as possible. If we now choose a certain lag time  $\tau$ , the model is exact for this specific choice, but predictions for longer times become inaccurate in the sense that (as mentioned in section 2.4.2) there is an error  $E(\tau)$  depending on the discretization and on the choice of the lag time  $\tau$ . We wish to check if

$$(\mathbf{T}(\tau))^k \approx \mathbf{T}(k\tau), \quad (2.6.6)$$

holds where the left side is the prediction of the model, while the right side is the exact, observed dynamics. For more practical purposes we compare the result for different initial distributions and evaluate the result directly from the estimated model as well as from the observed trajectory and compare for a range of lag times.

In section 2.4.2 we have formulated conditions for choosing a discretization and a lag time  $\tau$  that minimize the discretization error of a MSM. However, in practice it is essential to conduct a test whether a lag time and a discretization have been chosen so that the Markov model obtained is at least consistent with the data used to parametrize it within statistical error. In the previous discussion the discretization error was measured as the difference between Markov model propagation and true propagation in the continuous space. Since the exact propagation is usually not accessible, it is easier to measure the propagation error on the discrete space directly. In particular, we are interested in checking whether Eq. (2.6.6) with the estimated approximations

$$(\hat{\mathbf{T}}(\tau))^k \approx \hat{\mathbf{T}}(k\tau), \quad (2.6.7)$$

holds within statistical uncertainty. Here,  $\hat{\mathbf{T}}(\tau)$  is the transition matrix estimated from the data at lag time  $\tau$  (the Markov model), and  $\hat{\mathbf{T}}(k\tau)$  is the transition matrix estimated from the same data at longer lag times  $k\tau$ . Note that when the non-reversible maximum likelihood estimator, Eq. (2.6.4), is used, this approximation is trivially exact for  $k = 1$  since the Markov model was parametrized at lag time  $\tau$  to match exactly the observed transition probabilities. Also, for all

$k \gg t_2/\tau$ , the approximation should always be good as well, as Markov models correctly model the stationary distribution, even for bad choices of  $\tau$  and discretization (see section 2.4.2). Thus, this test is only sensitive in ranges of  $k$  greater than one and smaller than the global relaxation time of the system.

There are various ways how a test of Eq. (2.6.7) could be implemented, but it should consider the following points:

1. For large transition matrices, individual elements of  $\hat{\mathbf{T}}(k\tau)$  or  $(\hat{\mathbf{T}}(\tau))^k$  can be very uncertain, and comparing  $M \times M$  elements may be cumbersome. Therefore, we suggest to compare the probability of being in a given set of states,  $A$ , when starting from a well-defined starting distribution. This narrows the test down to few observables and allows to check the kinetics of states that are of special interest, such as folded / unfolded or metastable states.
2. The test should be done for all times  $k\tau$  for which trajectory data is available. Tests that compare Markov models that differ only one lag step ( $\tau$  and  $2\tau$ ) are likely to be unreliable as small approximation errors at short times may amplify at long times.
3. The quality of the approximation in Eq. (2.6.7) should be judged within the statistical uncertainties induced by the data.

Here we present an implementation that takes these properties into account [1]. Let  $\pi$  be the stationary probability of the Markov model  $\hat{\mathbf{T}}(\tau)$  and

$$\pi_i^{[A]} := \frac{\pi_i \mathbf{1}_A(i)}{\sum_{j \in M} \pi_j \mathbf{1}_A(j)}$$

the corresponding normalized stationary distribution restricted to an arbitrary set  $A$ . As a model test, the following “relaxation experiment” may be carried out: Using  $\pi^{[A]}$  as initial probability vector, the probability of being at the same set at all later times  $k\tau$  is computed according to the observed trajectory data and the Markov model, and is subsequently compared. The trajectory-based time-dependence of the probability to be in set  $A$  after time  $k\tau$  with starting distribution  $\pi(A)$  is

$$\hat{p}_{AA}^{\text{obs}}(k\tau) = \sum_{i,j \in M} \pi_i^{[A]} \hat{T}_{ij}^{\text{obs}}(k\tau) \mathbf{1}_A(i) \quad (2.6.8)$$

where  $\hat{\mathbf{T}}^{\text{obs}}(k\tau)$  is the trajectory-based estimate of the conditional jump probabilities. Likewise, the probability to jump to  $A$  according to the MSM is given by:

$$\hat{p}_{AA}^{\text{msm}}(k\tau) = \sum_{i,j \in M} \pi_i^{[A]} \mathbf{T}^k(\tau) \mathbf{1}_A(i). \quad (2.6.9)$$

Testing the validity of the Markov model then amounts to testing how well the equality

$$\hat{p}_{AA}^{\text{obs}}(k\tau) = \hat{p}_{AA}^{\text{msm}}(k\tau) \quad (2.6.10)$$

holds, which is essentially a test of the Chapman-Kolmogorov property. Note that the initial distribution  $\pi^{[A]}$  is simply a chosen reference distribution with respect to which the comparison is made.

The equality in Eq. (2.6.10) is not expected to hold exactly as a result of statistical uncertainties caused by the fact that only a finite number of transitions are available to estimate the true transition probabilities. To account for this, we use the variance in the likelihood computed for the Dirichlet distribution of a two state model with the two states  $\{A, \Omega \setminus A\}$  as derived in section 2.6.4. The uncertainties  $\sigma_{AA}^{\text{obs}}(k\tau)$  (one-sigma standard error) of the transition probabilities is then estimated as

$$\sigma_{AA}^{\text{obs}}(k\tau) = \sqrt{k \frac{\hat{p}_{AA}^{\text{obs}}(k\tau)(1 - \hat{p}_{AA}^{\text{obs}}(k\tau))}{\sum_{i \in A} \hat{z}_i(k\tau)}}$$

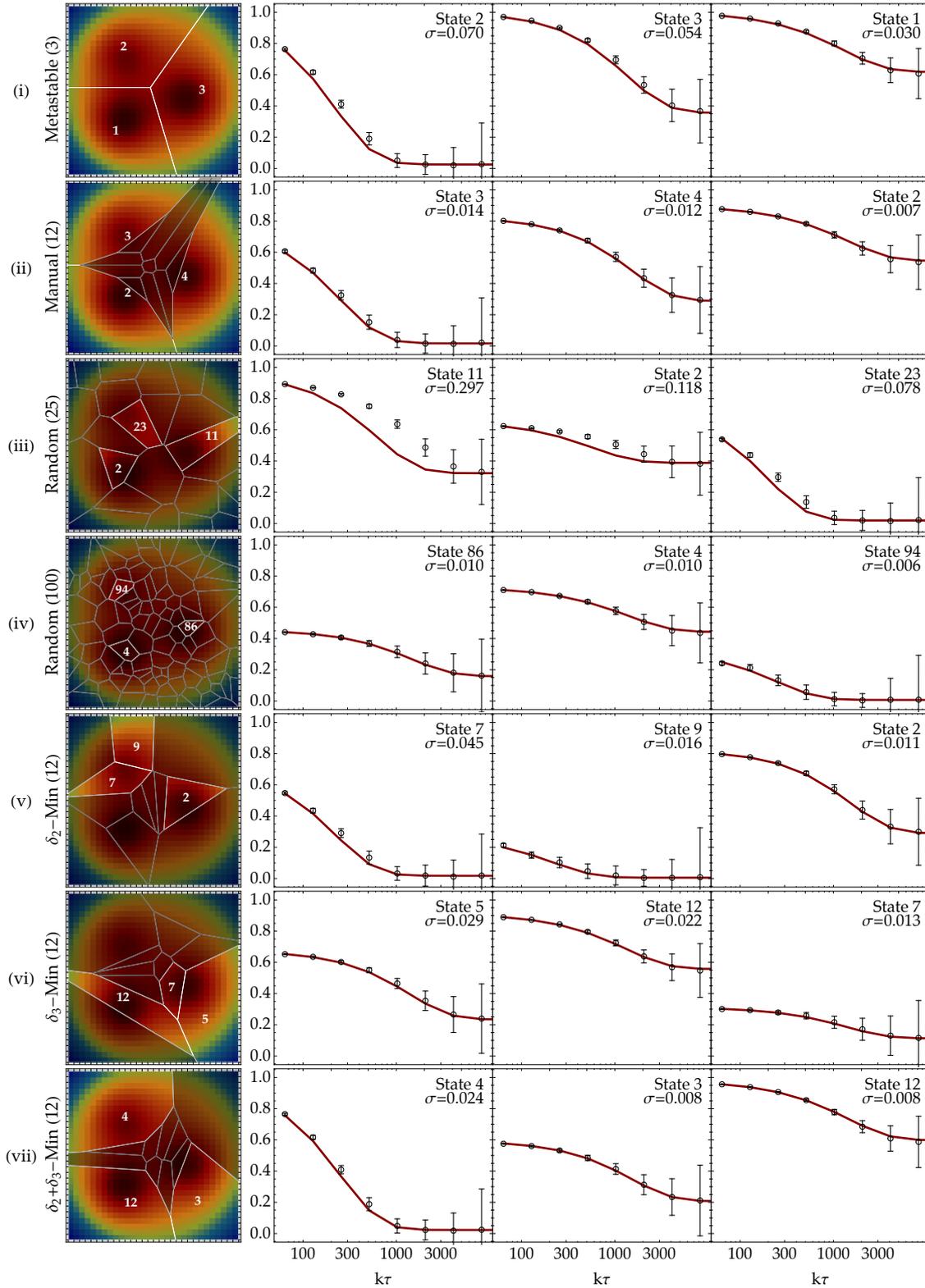
with  $\hat{z}_i(k\tau)$  being the number of transitions originating in state  $i$  while assuming a lag time of  $k\tau$ . The test finally consists of assessing whether Eq. 2.6.10 holds within these uncertainties.

For an illustration, a single 250,000 step trajectory was simulated in the 2D-three-well diffusion model (see appendix D for model details) started from the energy minimum at  $x = (10, 10)$ . Figure 2.5 shows the corresponding results: For each of the seven different discretization shown in the first column of Figure 2.5 the probability to stay in a state is shown for the three states with the largest Markov model error (highlighted in 2.5, left column). It is apparent that the metastable three-state discretization (Figure 2.5(i)) performs well, however sacrificing metastability in order to more finely discretize the transition region, generating a superior discretization (Figure 2.5(ii) and (v)-(vii)). The ‘‘uniform’’ random 25-state clustering (Figure 2.5(iii)) performs worst but can be improved significantly by using more states (Figure 2.5(iv)). This further supports our theoretical finding that either a clustering adapted to the eigenfunctions or using more states can improve the quality of the constructed MSM.

## 2.7 DISCUSSION AND CONCLUSION

Markov modeling is a simulation analysis tool which is rapidly gaining popularity in the MD community. In this chapter we have summarized the state-of-the-art of generation and validation of Markov models of molecular kinetics. The theoretical framework and important concepts have been laid out and, in addition, some important methodological gaps have been filled.

As shown in section 2.1, any physically reasonable implementation of equilibrium molecular dynamics can be understood in terms of



**Figure 2.5** – Chapman-Kolmogorov test for the three-well-diffusion-model (see also Figure 2.4). For each of seven discretizations (first column), the Chapman-Kolmogorov test is shown for the three states with the largest error (labeled with white numbers in the first column). Relaxation curves from a 250,000 step trajectory,  $\hat{p}_{AA}^{\text{obs}}(k\tau)$  (black) along with the uncertainties  $\sigma_{AA}^{\text{obs}}(k\tau)$  are compared to the model prediction,  $\hat{p}_{AA}^{\text{msm}}(k\tau)$  (red). The total error  $\sigma$  given in the top right corners is measured as the 2-norm of the vector containing the differences  $\hat{p}_{AA}^{\text{obs}}(k\tau) - \hat{p}_{AA}^{\text{msm}}(k\tau)$  for time points in the range  $k \in \{1, 2, 4, 8, 16, 32, 64, 128\}$  and a chosen lagtime of  $\tau = 64$ .

relaxation processes that are described by the eigenfunctions of the dynamical operator, i.e. the transfer operator  $T_\tau$ . The role of these eigenfunctions in molecular kinetics cannot be overemphasized, irrespective of whether Markov models are used or not, for they unambiguously yield a structural dynamical interpretation of the relaxation processes. Each eigenfunction is linked to one eigenvalue with a corresponding relaxation time scale that is accessible experimentally, thus Markov models can serve as a means to interpret kinetic experimental data (we will address this idea again in chapter 5). From a modeling point of view, the dynamical decomposition in Eq. (2.2.8) shows that these eigenfunctions define coordinates in which slow and fast dynamics can be separated exactly. Indeed, they are the only choice of coordinates for which such a separation is possible and any different attempt to model the dynamics *via* a projection onto slow degrees of freedom or order parameters will necessarily introduce memory terms that are challenging to deal with [91]. In chapter 5 we will introduce a new way to circumvent the projection problem by relaxing the constraints of a Markov model.

One of the key insights from the work in this chapter that was published in Ref. [1] is that the discretization error made by using a Markov model on a discrete state space can be controlled by choosing the discretization and the lag time adequately (see section 2.4). In particular, the quality of the Markov model depends on how well the discretization approximates the slowly relaxing eigenfunctions of the true dynamics. It is shown in section 2.4.2 how the Markov model can be used to precisely approximate only selected slow processes with relatively few discrete states slicing the state space finely in regions where the corresponding eigenfunctions change rapidly while leaving the discretization coarse in regions where only the fast eigenfunctions vary. This answers a key concern about discretization-based kinetic model approaches, namely that for complex macromolecular systems there is no hope to enumerate all energy basins in the discrete model. The present analysis shows that this is indeed not necessary and that in principle, very few states are sufficient to obtain an excellent model. Moreover, the analysis also shows that metastable partitions suggested in previous works [52, 51] are good among partitions where the number of states  $n$  is allowed to be less or equal to the number of metastable states in the system, but that the approximation error can be further reduced by increasing the number of partitions, even if this means that the individual discrete states are no longer metastable.

This immediately raises the question how such a discretization can be created for a complex molecular system where the true eigenfunctions are initially unknown and this issue has not yet been solved. Based on current results, it is clear that subdividing discrete states should always reduce the discretization error. Thus, when geometric

clustering methods are used to subdivide state space, it is advisable to use as many clusters as possible without running into serious statistical problems. In the longer term, much better discretization can be expected from methods that adaptively discretize in an iterative manner. For example, first an initial Markov model is created based on a geometric clustering, these clusters are then subdivided, providing a finer Markov model. The discretization error of the coarser model with respect to the finer model is computed using the error bound from section 2.4.2, and it is then decided which states are kept, lumped or split. An adaptive method based on maximizing metastability has been proposed in [52], and a similar approach may be followed by minimizing the error bound from section 2.4.2 as we have qualitatively demonstrated.

In a broader sense, adaptive space discretization methods based on error bounds are commonly and successfully used in other disciplines where equations must be solved on a grid, e.g., in fluid mechanics and engineering. Moving to such approaches, MD becomes more and more a numerical analysis problem of molecular phase spaces, and may therefore benefit from the understanding of discretization methods that have been acquired in scientific computing.

# 3

---

## EFFICIENT COMMITTOR COMPUTATION

---

This chapter is based on the publication

- [2] Prinz, J.-H., Held, M., Smith, J. C. & Noé, F. Efficient Computation, Sensitivity, and Error Analysis of Commitor Probabilities for Complex Dynamical Processes. *Multiscale Model. Sim.* **9**, 545–567 (2011). doi [10.1137/100789191](https://doi.org/10.1137/100789191).

### 3.1 INTRODUCTION

In many cases, characterizing the dynamics between two subsets  $A, B \subset \Omega$  of configurational space  $\Omega$  provides a satisfactory picture of the process (e.g. in protein folding dynamics,  $A$  being unfolded and  $B$  native [35]), whereas in other cases the simultaneous consideration of multiple substates is necessary. It is now widely recognized that the committor probability, also called splitting probability or probability of folding in some contexts, is one of the central mathematical objects needed for intersubstate processes to be characterized [92, 93, 94, 95, 96, 97, 98, 99].

The *committer*  $q$  is a state function that provides the probability that a system initially at state  $x \in \Omega$  will encounter  $B$  first rather than  $A$  under the action of the given system dynamics. The committor thus defines a dynamical reaction coordinate, which has the advantage over *ad hoc* reaction coordinates that it does not bring the danger of concealing relevant dynamics of the system. In this chapter, we investigate how the committor probability can be efficiently computed for large-scale systems and study its sensitivity as well as its uncertainty in cases where the full dynamics has been inferred from a finite set of observations.

## 3.2 THEORY

## 3.2.1 Definition

Given a time-discrete Markov process with state space  $\Omega$  as defined in chapter 2, the committor values

$$q : x \in \Omega \mapsto q(x) \in [0, 1]$$

for each state  $x \in \Omega$  are defined as the probability that a process starting in state  $x$  will reach subset  $B \subset \Omega$  before it hits subset  $A \subset \Omega$ . Thus it can simply be evaluated by running Monte Carlo simulations starting in  $x$  until they reach either  $A$  or  $B$  with the committor being the fraction of simulations that reach  $B$  first. For the examined time-discrete case we also need to specify the point in time, when we test if a subset is hit. A check *before* the first transition means that starting in one of the two sets is already a hit and so we define naturally the committor values for these states  $x \in A$  and  $x \in B$  to be

$$\begin{aligned} q(x) &= 0, \forall x \in A \\ q(x) &= 1, \forall x \in B \end{aligned}$$

whereas checking *after* the first transition allows for a non-zero probability to leave the two subsets and henceforth leads to slightly different boundary conditions for the committor values

$$\begin{aligned} q(x) &\geq 0, \forall x \in A \\ q(x) &\leq 1, \forall x \in B. \end{aligned}$$

## 3.2.2 Relevance of the committor

Given a dynamical model, let us examine a number of aspects of the system dynamics that can be accessed *via* the committor probability:

Firstly, for a continuous state space  $\Omega$  all subsets  $I \subset \Omega$  of constant committor probability

$$I : q \in [0, 1] \mapsto I(q) := \{x \in \Omega \mid q(x) = q\} \subset \Omega$$

part the state space into two disjoint subsets

$$A \subset \Gamma(q) := \{x \in \Omega \mid q(x) < q\} \subseteq \Omega, q > 0$$

and

$$B \subseteq \Gamma^+(q) := \{x \in \Omega \mid q(x) > q\} \subseteq \Omega, q < 1.$$

The committor is thus a measure for the progress of a process or reaction, i.e. it is the ideal reaction coordinate for the process  $A \rightarrow B$  [92, 96, 99]. Of special interest in this context is the isocommittor surface  $\Omega^{\text{TS}} = I(0.5) \subset \Omega$ , which can be interpreted as the transition

state ensemble in protein folding theory [100]. If the state space is discrete or infinite but without topological structure the definition of  $I$ ,  $I^+$  and  $I^-$  has to be seen in a less strict sense. The transition state ensemble is then  $\Omega^{\text{TS}} = \bigcup_{q \approx 0.5} I(q) \subset \Omega$ , the set of states of approximately committor value 0.5.

Once the committor has been computed, the change of any state variable (an observable)

$$a : x \in \Omega \mapsto a(x) \in \mathbb{R}$$

may be monitored along the  $A \rightarrow B$  coordinate using

$$a(q) = \mathbb{E}[a(x)]_{I(q)} = \frac{\int_{I(q)} dx \pi(x) a(x)}{\int_{I(q)} dx \pi(x)},$$

with  $\pi$  being the stationary distribution. If the stationary distribution is unique (see chapter 2.2), one can define a Potential of Mean Force (PMF)  $V$  given by

$$V(q) = V(q_0) - \beta^{-1} \left( \log \frac{\int_{I(q)} dx \pi(x)}{\int_{\Omega} dx \pi(x)} - \log \frac{\int_{I(q_0)} dx \pi(x)}{\int_{\Omega} dx \pi(x)} \right) \quad (3.2.1)$$

which can be interpreted as the origin of a virtual force causing changes in the stationary probability along the committor value  $q$  at inverse temperature  $\beta$ . The potential  $V(q_0)$  at some arbitrary committor value  $q_0$  can be chosen freely to fix the absolute height of the potential function. For discrete state spaces  $\Omega$  this definition fails and one has to introduce e. g. binning.

For simplicity, we assume next, that the system dynamics can be given in terms of a transition matrix  $\mathbf{T}$  or a rate matrix  $\mathbf{K}$  over the countable state space  $\Omega$  to compute transport properties from  $A$  to  $B$  via Transition Path Theory (TPT) [101, 63]. In particular, the reactive flux  $F$  between two states  $i$  and  $j$  is then given by

$$F_{ij} = \pi_i q_i^- K_{ij} q_j^+$$

for rate matrices [63], or

$$F_{ij}(\tau) = \pi_i q_i^- T_{ij}(\tau) q_j^+$$

if the transition probability matrix  $\mathbf{T}$  is used [35]. The forward  $q^+$  and backward  $q^-$  committor will be defined later in section 3.2.3. The reactive flux  $F_{ij}$  is proportional to the probability that the transition  $i \rightarrow j$  is part of the reactive trajectories from  $A$  to  $B$ . The reactive trajectories directly go from  $A$  to  $B$  (i.e. never hitting  $A$  or  $B$  except from the beginning or end) and thus effectively contribute to the  $A \rightarrow B$  reaction, whereas trajectories that start in  $A$  and directly go back to  $A$  are not reactive. Also the net transport through  $i \rightarrow j$  can be expressed using the reactive flux by

$$F_{ij}^+ = \max\{F_{ij} - F_{ji}, 0\}, \quad (3.2.2)$$

which defines a network-flow out of A and into B that can be decomposed into a set of  $A \rightarrow B$  reaction pathways along with their respective probabilities [63, 35, 101].

Lastly, one can also express different rates using the committor probability. The *global*  $A \rightleftharpoons B$  flux  $F$  from TPT is defined as the average number of trajectories traveling between A and B per time unit. It can be expressed by [63]

$$F = \sum_{i \in A, j \notin A} F_{ij} = \sum_{i \in A, j \notin A} F_{ij}^+ \quad (3.2.3)$$

and is the inverse expected time needed for an  $A \rightleftharpoons B$  cycle. It can be composed into the inverse forward rate constant  $k_{AB}$  and the inverse backward rate constant  $k_{BA}$  by

$$F^{-1} = k_{AB}^{-1} + k_{BA}^{-1}.$$

In order to calculate these rate constants directly we need the *milestoning* probability  $\pi_A^{\text{mile}}$ , i.e. the equilibrium expectation over all states  $x \in \Omega$  that the process has last been in A while  $\pi_B^{\text{mile}} = 1 - \pi_A^{\text{mile}}$  is the expectation that it has been in B before. The milestoning probability can be computed by

$$\pi_A^{\text{mile}} \equiv \sum_{j \in \Omega} \pi_j q_j^-$$

and using a detailed balance condition

$$\pi_A^{\text{mile}} k_{AB} = (1 - \pi_B^{\text{mile}}) k_{BA},$$

we can derive the expression for the  $A \rightarrow B$  rate constant  $k_{AB}$  as

$$k_{AB} = \left( \pi_A^{\text{mile}} \right)^{-1} F.$$

For alternative rate definitions and their relations see e.g. [102].

Given the fundamental relevance of the committor probability in the characterization of dynamical processes, it is important to be able to compute  $\mathbf{q}$  efficiently and to understand its sensitivity to perturbations; especially in cases where the system dynamics can be computed only approximately, e.g., by some sampling scheme such as molecular dynamics simulations or Monte-Carlo dynamics.

### 3.2.3 Committor Equations

We will now give a derivation of the well-known committor equation using a recursive approach for the case of a finite state space  $\Omega$  with  $|\Omega| = M$  states. For this we will construct a set of linear equations, that uniquely define the committor and only use assumptions which are in accordance with the definition of the committor. To be precise, given a transition matrix  $\mathbf{T} \in \mathbb{R}^{M \times M}$  we define the *forward committor*  $\mathbf{q}^+$  to be the (unique) solution of the following set of equations:

1. The committor for states  $a \in A$  are zero by definition

$$q_x^+ \equiv 0, \forall x \in A,$$

2. for states  $b \in B$  the committor is one

$$q_x^+ \equiv 1, \forall x \in B,$$

3. while for all intermediate states  $i \notin A \cup B$  the committor is chosen in accordance with the dynamics given in the transition matrix  $\mathbf{T}$ : That means that the committor  $q_i$  of state  $i$  can be set into relation to the committor values of all connected states which need to be weighted with the conditional probability to reach this particular neighboring state. This mathematically be expressed by

$$q_i^+ = \sum_{j \in \Omega} T_{ij} q_j^+, \quad i \notin A \cup B.$$

These equations can conveniently be reformulated in matrix notation as the solution of the set of linear equations

$$(\mathbf{T} - \text{Id})\mathbf{q} = 0 \tag{3.2.4}$$

subject to boundary conditions

$$q_{x \in A} = 0, \quad q_{x \in B} = 1.$$

In contrast to this *forward* definition, a backwards directed committor probability,  $q_i^-$  is defined respectively as the probability that being in state  $i$ , the system was in B last rather than in A. In order to obtain the backward committor, we use the previously defined backward propagator

$$\mathbf{T}^- := \mathbf{\Pi}^{-1} \mathbf{T}^T \mathbf{\Pi},$$

which contains the probabilities that if the system is in state  $i$  then it came from state  $j$  (see section 2.2). Proceeding in analogy to the forward committor  $\mathbf{q}^+$  we get

$$\begin{aligned} q_i^- &= 0 & \text{if } i \in A \\ q_i^- &= 1 & \text{if } i \in B \\ \sum_{j \in \Omega} T_{ij}^- q_j^- &= 0 & \text{if } i \notin A \cup B. \end{aligned}$$

for the backward committor  $\mathbf{q}^-$ . For reversible dynamics the forward and backward propagations are equal (see section 2.2) from which it immediately follows that

$$q_i^- = 1 - q_i^+$$

holds. One should note, that the different definitions can lead to ambiguity when speaking just about the committor. One has to specify:

1. the target set A or B to be hit first,
2. the direction of propagation: *forward* ( $\mathbf{T}$ ) or *backward* ( $\mathbf{T}^-$ ),
3. and for time-discrete processes, when to measure the hitting:  
At time step  $t$  or the next step  $t + 1$ .

This, in principle, allows for 8 different committor functions and if not stated otherwise, we denote with  $q$  the forward committor to hit set B at time  $t$ .

### 3.2.4 Rate Matrix

Given the rate matrix  $\mathbf{K} \in \mathbb{R}^{M \times M}$ , we can use a similar argument as for the time discrete case and derive expressions for the committor

$$\begin{aligned} q_i &= 0 & \text{if } i \in A \\ q_i &= 1 & \text{if } i \in B \\ \sum_{j \in \Omega} K_{ij} q_j &= 0 & \text{if } i \notin A \cup B. \end{aligned}$$

and the corresponding equations hold also for the backward committor. A proof can be found in [103, 63].

### 3.2.5 Transforming between Rate and Transition Matrices

It turns out that there is a simple way to transform rate matrices into transition matrices and *vice versa* that leaves the committor probabilities unchanged. This transformation is useful when a method is available to compute the committor from the transition matrices, but not for rate matrices, or *vice versa*.

**Theorem 1.** Let  $\mathbf{K} \in \mathbb{R}^{M \times M}$  be a rate matrix and  $\mathbf{T}_c(\mathbf{K}) \in \mathbb{R}^{M \times M}$  a stochastic matrix defined by the transformation

$$\mathbf{T}_c(\mathbf{K}) = c \mathbf{K} + \text{Id}, \quad 0 < c < c^{\max} \quad (3.2.5)$$

with  $c^{\max} = -\min_{i \in \Omega} K_{ii}$  representing the largest entry in magnitude in the rate matrix. Then  $\mathbf{T}(\mathbf{K})$  and  $\mathbf{K}$  have the same committor probabilities for any choice of  $A, B \subset \Omega$ .

*Proof.* For arbitrary A and B we need to show, that if  $\mathbf{K}$  fulfills the committor equations for rate matrices, then  $\mathbf{T}_c(\mathbf{K})$  has to fulfill the committor equation for transition matrices. By definition this is true for all states  $x \in I \equiv A \cup B$ . For all other states we start with the committor equation for transition matrices and replace the transition matrix by  $\mathbf{T}_c(\mathbf{K})$  and get

$$\sum_{j \in I} (c \cdot K_{ij} + \delta_{ij}) q_j = q_i$$

which can be simplified to the committor equation for rate matrices

$$c \cdot \sum_{j \in I} K_{ij} q_j = 0.$$

The theorem is independent of the choice of  $c$ , but  $0 < c < c^{\max}$  assures that the row sum of zero in the rate matrix translates into a row-stochastic transition matrix  $\mathbf{T}_c(\mathbf{K})$ . A more direct proof is the fact that scaling of matrices by a constant factor and adding multiples of the identity matrix do not change the eigenvectors of a matrix. Thus  $\mathbf{T}_c(\mathbf{K})$  inherits the same eigenvectors as  $\mathbf{K}$ , but with different eigenvalues. Since we show later, that aside the boundary conditions, the committor can be computed from the eigenvectors, both matrices will result in the same committor probabilities. However, it is important to note that  $\mathbf{T}_c(\mathbf{K})$  will not reproduce the dynamical behavior of the rate matrix  $\mathbf{K}$  on any but infinite timescales.  $\square$

The theorem allows the methods explained later to be used with equal computational effort to rate matrices. In particular, the theorem will keep a potentially sparse structure of the rate matrix and allows to use sparse eigensystem algorithms in these cases as well. As a side remark the committor equations for the rate matrix takes the particular simple form

$$\mathbf{K}\mathbf{q} = 0, \quad q_{x \in A} = 0, \quad q_{x \in B} = 1.$$

The committor equations in Eq. (3.2.4) can easily be solved with any linear systems solver. When the system is very large and sparse, a sparse linear systems solver may still be able to handle them efficiently. An alternative approach to compute the committor probability from  $\mathbf{K}$  has been proposed in [104]. However, this approach requires the  $\mathbf{K}$ -matrix to be inverted, which effectively limits its applicability to systems of  $|\Omega| \leq 10^4$  states.

### 3.3 ALTERNATIVE COMMITTOR COMPUTATION

An alternative view is obtained when formulating the committor problem in terms of the dominant eigenvectors of either  $\mathbf{K}$  or  $\mathbf{T}(\tau)$ . This is useful from a numerical point of view, because efficient solvers, such as the Power method or Krylov subspace methods, exist for dominant Eigenvectors. Moreover, it is useful from a physical standpoint as it allows the committor to be understood in terms of the slowest relaxation process of the system.

An approach to approximate  $\mathbf{q}$  in terms of the second eigenvector of  $\mathbf{K}$  or  $\mathbf{T}(\tau)$  has been proposed in [105]. This approach is valid only if the second eigenvector is similar to the  $A \rightarrow B$  committor and the second and third eigenvalues are well separated. In molecular processes, this is often referred to as a “two-state” process, where

there exists one slow process that is clearly separated from all other processes in terms of timescales ( $\lambda_3 \ll \lambda_2 \approx 1$ ). In the following, we will derive equations that allow the committors to be computed exactly in terms of dominant eigenvectors for any Markovian system.

### 3.3.1 $A \rightarrow B$ Committor

We construct an altered transition matrix  $\tilde{\mathbf{T}}$  with absorbing states A and B from the original dynamics in  $\mathbf{T}$  by

$$T^{\text{AB}}_{ij} = \begin{cases} T_{ij} & \text{if } i \notin A \cup B \\ 1 & \text{if } i \in A \cup B, j = i \\ 0 & \text{if } i \in A \cup B, j \neq i \end{cases}$$

assuming that the underlying dynamics is ergodic or equivalently  $\mathbf{T}$  irreducible, positive recurrent and aperiodic. We then define a transition matrix for infinite times  $\tilde{\mathbf{T}}^\infty$  that transports any initial distribution infinitely into the future

$$\tilde{\mathbf{T}}^\infty = \lim_{k \rightarrow \infty} \tilde{\mathbf{T}}^k \quad (3.3.1)$$

and consequently directly into either A or B which follows from ergodicity. Given an ensemble which is located completely at state  $x \in \Omega$ , its committor value  $q_x$  is then given by the total probability to find this ensemble in set B after it has been propagated infinitely into the future using  $\tilde{\mathbf{T}}^\infty$ . The initial ensemble can be expressed by a canonical unit vector  $\mathbf{e}_i : e_{ix} = \delta_{ix}$  leading to

$$q_x = \sum_{b \in B} (\mathbf{e}_x^T \tilde{\mathbf{T}}^\infty)_b = \sum_{b \in B} \tilde{T}^\infty_{xb} \quad (3.3.2)$$

as a way to compute the committor. In the following we will show that  $\tilde{\mathbf{T}}^\infty$  and thus the committor  $\mathbf{q}$  are computationally fast and robust to derive in terms of the eigenvectors of  $\mathbf{T}$ . Without loss of generality we treat here the case where we reduce the sets  $A = \{a\}$  and  $B = \{b\}$  to a single state each.

In cases where the sets are larger they can simply be aggregated into a single state in the definition of  $\tilde{\mathbf{T}}$ . For this one defines a crisp membership matrix

$$\chi \in [0, 1]^{M' \times M} : i \in \Omega^X, x \in \Omega \mapsto [0, 1]$$

with a reduced state space  $\Omega^X \equiv \Omega \setminus (A \cup B) \cup \{a, b\}$  and  $M' = |\Omega^X|$  that simply clusters the states into a single state by defining

$$\chi_{ix} \equiv \chi_i(x) = \begin{cases} 1 & \text{if } i = a \wedge x \in A \\ 1 & \text{if } i = b \wedge x \in B \\ 1 & \text{if } i \notin \{a, b\} \wedge i = x \\ 0 & \text{else} \end{cases}.$$

Now Eqs. 2.4.3 and 2.4.4 can be used to get the clustered transition matrix  $\mathbf{T}^X$  which can in this case be written as

$$\mathbf{T}^X = (\chi^T \Pi \chi)^{-1} \chi^T \Pi \mathbf{T} \chi$$

in matrix notation. Now we can proceed as before using the clustered transition matrix and construct  $\tilde{\mathbf{T}}$ .

When diagonalizing  $\tilde{\mathbf{T}}$ , one obtains

$$\begin{aligned} \tilde{\mathbf{T}}^\infty &= \lim_{k \rightarrow \infty} \mathbf{R} \cdot \text{diag}(\lambda_1^k, \dots, \lambda_N^k) \cdot \mathbf{R}^{-1} \\ &= \mathbf{R} \cdot \left( \lim_{k \rightarrow \infty} \text{diag}(\lambda_1^k, \dots, \lambda_N^k) \right) \cdot \mathbf{R}^{-1}, \end{aligned} \quad (3.3.3)$$

with  $\mathbf{R}$  being the matrix of right eigenvectors of the infinite-time and A,B-absorbing transition matrix  $\tilde{\mathbf{T}}^\infty$  and  $\lambda_i$  the corresponding eigenvalues, which are sorted as usual from the largest to the smallest absolute value. It follows from the Perron-Frobenius theorem that there exist exactly two eigenvectors with eigenvalue one while the absolute value of all other eigenvalues are strictly smaller than one. Since,

$$\lim_{k \rightarrow \infty} |\lambda_i^k| = 0, \forall \lambda_i < 1,$$

we get the simplified representation

$$\tilde{\mathbf{T}}^\infty = \mathbf{R} \cdot \text{diag}(1, 1, 0, \dots, 0) \cdot \mathbf{R}^{-1}. \quad (3.3.4)$$

Using, that the right eigenvectors can be expressed in terms of the left ones if these are chosen to be  $\mathbf{L}^T := \mathbf{R}^{-1}$ , allows to reduce the computation to

$$\begin{aligned} \tilde{\mathbf{T}}^\infty &= \mathbf{R} \cdot \text{diag}(1, 1, 0, \dots, 0) \cdot \mathbf{L}^T \\ &= \mathbf{R}_{[1:2,:]} \mathbf{L}_{[1:2,:]}^T. \end{aligned} \quad (3.3.5)$$

This means that once we have the basis of left eigenvectors that equal  $\mathbf{R}^{-1}$  we can avoid the expensive calculation of the inverse. Although this is no advantage in general, in the present case, the left eigenvectors of  $\tilde{\mathbf{T}}^\infty$  take a particularly simple form:  $\tilde{\mathbf{T}}^\infty$  will transported any initial ensemble only to A or B so that any row in  $\tilde{\mathbf{T}}^\infty$  different from  $a$  or  $b$  has to be zero. This structure of  $\tilde{\mathbf{T}}^\infty$  implies that the left eigenvectors to the eigenvalue of one are a linear combination of the canonical unit vectors  $\mathbf{e}_a$  and  $\mathbf{e}_b$  which leads to

$$\begin{aligned} \mathbf{L}_{[1:2,:]}^T &= \mathbf{S} \cdot [\mathbf{e}_a, \mathbf{e}_b]^T \\ &= \mathbf{S} \cdot \text{Id}_{\{a,b\},\cdot} \end{aligned}$$

using a mixing matrix  $\mathbf{S} \in \mathbb{R}^{2 \times 2}$  and recall that  $\text{Id}_{\{a,b\},\cdot} \in \{0, 1\}^{2 \times M}$  contains only a one at column  $a$  in the first row and a second one in column  $b$  in the second row. Exploiting the fact that  $\tilde{\mathbf{T}}^\infty$  is still a stochastic matrix we can choose one of the right eigenvectors to be

the constant right Perron Eigenvector,  $\mathbf{1} := \mathbf{R}_{[1,\cdot]} = (1, \dots, 1)$ . Conclusively, only one (non-constant) right eigenvector  $\mathbf{r} := \mathbf{R}_{[2,\cdot]}$  needs to be known for

$$\tilde{\mathbf{T}}^\infty = [\mathbf{1}, \mathbf{r}] \cdot \mathbf{S} \cdot \text{Id}_{[\{a,b\},\cdot]}, \quad (3.3.6)$$

to be computed. Using Eq. (3.3.2) we find the compact form

$$\begin{aligned} [\mathbf{q}^{\rightarrow A}, \mathbf{q}^{\rightarrow B}] &= \tilde{\mathbf{T}}^\infty \cdot \text{Id}_{[\cdot, \{a,b\}]} \\ &= [\mathbf{1}, \mathbf{r}] \cdot \mathbf{S} \cdot \text{Id}_{[\{a,b\},\cdot]} \cdot \text{Id}_{[\cdot, \{a,b\}]} \\ &= [\mathbf{1}, \mathbf{r}] \cdot \mathbf{S} \end{aligned} \quad (3.3.7)$$

for the committor  $\mathbf{q}^{\rightarrow B}$  and  $\mathbf{q}^{\rightarrow A}$  respectively where we used that  $\text{Id}_{[\cdot, \{a,b\}]}$  selects columns  $a$  and  $b$ . Thus the committor is a linear combination of the two right Perron eigenvectors of  $\tilde{\mathbf{T}}^\infty$ . To derive an expression for the mixing matrix  $\mathbf{S}$  we make use of the boundary conditions that

$$\begin{aligned} q_A^{\rightarrow A} &= 1 & q_A^{\rightarrow B} &= 0 \\ q_B^{\rightarrow A} &= 0 & q_B^{\rightarrow B} &= 1 \end{aligned}$$

is true by definition. Multiplying Eq. (3.3.7) with  $\text{Id}_{[\{a,b\},\cdot]}$  from the left will select rows  $a$  and  $b$  results in

$$\begin{aligned} \text{Id}_{[\{a,b\},\cdot]}[\mathbf{q}^{\rightarrow A}, \mathbf{q}^{\rightarrow B}] &= \text{Id}_{[\{a,b\},\cdot]} \cdot [\mathbf{1}, \mathbf{r}] \cdot \mathbf{S} \\ \begin{pmatrix} q_a^{\rightarrow A} & q_a^{\rightarrow B} \\ q_b^{\rightarrow A} & q_b^{\rightarrow B} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & r_a \\ 1 & r_b \end{pmatrix} \cdot \mathbf{S} \end{aligned}$$

with the simple solution

$$\mathbf{S} = \begin{pmatrix} 1 & r_a \\ 1 & r_b \end{pmatrix}^{-1}. \quad (3.3.8)$$

Using Eq. (3.3.8) in (3.3.7) yields the final solution

$$[\mathbf{q}^{\rightarrow A}, \mathbf{q}^{\rightarrow B}] = [\mathbf{1}, \mathbf{r}] \cdot \begin{pmatrix} 1 & r_a \\ 1 & r_b \end{pmatrix}^{-1}$$

or expressed for the single committor elements  $q_x$  by

$$q_x^{\rightarrow B} = \frac{r_x - r_a}{r_b - r_a}. \quad (3.3.9)$$

Finally, we have avoided the inversion of the matrix  $\mathbf{R}$  of right eigenvectors required in Eq. (3.3.4) and instead reduced the effort to computing one largest non-trivial right eigenvector. This is consistent with the computational effort of solving the system of linear equation in Eq. (3.2.4).

Based on Eq. (3.3.9), the committor probability can be computed for large sparse transition matrices using e.g. the Power method [106], or, if the system dynamics is specified in terms of the rate matrix, this

computation can be performed by first applying the transformation (3.2.5). In the case when the Power method is used to solve for  $\mathbf{r}_2$ , the parameter  $c$  has to be larger than zero and strictly smaller than one, otherwise it cannot be proven that all eigenvalues, except one (the Perron-eigenvalue  $\lambda_1 = 1$ ), are inside the unit circle of the complex plane ( $|\lambda| < 1$ ), assuring convergence to the correct eigenvector. Since in large systems the eigenvalues are expensive to compute, a good guess is to choose  $c$  close to one, which, in most cases, will maximize the relative gap between the Perron-Eigenvalues and the next smaller eigenvalues and thus the rate of convergence. The advantage of the Power method is its simplicity, low memory requirement and applicability to sparse matrices, which allows it to treat very large systems ( $M \sim 10^6$ ) as shown later in the 3D model. However, in the case of slow processes indicated by a very small spectral gap, the Power method might converge too slowly or not at all due to numerical issues. In these cases more advanced methods like Krylov subspace methods (Arnoldi, Lanczos) [106] or graph theoretical approaches [107] can be a solution. One should note, that if the committor is not needed for the full state space, it is enough to compute the values of the eigenvector at A, B and the region of interest. This might be another possibility for speed up.

### 3.4 GENERALIZATION TO MULTIPLE STATES

In many applications, it is desirable to compute the committing probability between more than just two subsets. Consider a system for which a number  $N \geq 2$  of so called *core sets*

$$\mathcal{Y} = \{Y_1, \dots, Y_N\}, \quad Y_i \subset \Omega$$

have been defined. For each state  $x \in \Omega$  we want to evaluate the probability that the system dynamics will hit a specific core  $Y_i$  rather than any other core set  $\bigcup_{j \neq i} Y_j$  which resembles the  $A \rightarrow B$  committor in the two-state case  $N = 2$ . For simplicity we again assume single state cores  $Y_i = \{y_i\}$  and define a combined set  $Y$  containing all cores by

$$Y \equiv \bigcup_{i=1}^N Y_i = \{y_1, \dots, y_M\}.$$

The *committor matrix*

$$\mathbf{Q} \in [0, 1]^{M \times N} : x \in \Omega, n \in \{1, \dots, N\} \mapsto Q_{nx}$$

contains the  $N$  multi-committor probabilities and is defined by

$$\mathbf{Q} \equiv [\mathbf{q}^{\rightarrow Y_1}, \dots, \mathbf{q}^{\rightarrow Y_N}]$$

where the elements  $Q_{xn}$  contain the committor probability of encountering core  $n$  before any other core given the system was initially in state. Each row sums up to one and is strictly non-negative

$$Q_{xn} \geq 0, \quad \sum_{n=1}^N Q_{xn} = 1$$

so that  $\mathbf{Q}$  forms a fuzzy membership similar to the continuous memberships  $\chi$  as defined in section 2.4.1.

To solve this general case, all states in  $\mathcal{Y}$  are considered absorbing in the altered transition matrix, and a basis for all  $N$  eigenvectors to the eigenvalue of one is computed. Proceeding in analogy to the 2-state case the solutions are a linear combination of these  $N$  right eigenvectors. The parameters for the eigenvectors can then be computed using a simple matrix inversion in analogy to the two state case by

$$\mathbf{Q} = [\mathbf{R}_{[1,\cdot]}, \dots, \mathbf{R}_{[N,\cdot]}] \cdot \begin{pmatrix} \mathbf{R}_{[1,y_1]} & \cdots & \mathbf{R}_{[N,y_1]} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{[1,y_N]} & \cdots & \mathbf{R}_{[N,y_N]} \end{pmatrix}^{-1} \quad (3.4.1)$$

or in short

$$\mathbf{Q} = \mathbf{R}_{[1:N,\cdot]} \cdot \mathbf{R}_{[1:N,\mathcal{Y}]}^{-1}.$$

One eigenvector can be always chosen as the right constant Perron-Eigenvector, although this is for most eigenvector solver no advantage. The usage of a multi-state committor and its relation to memberships and projected dynamics is exploited extensively in milestone theory [108] and Multi-State Transition Interface Sampling (MSTIS) [109].

### 3.5 SENSITIVITY AND UNCERTAINTY

For a sensitivity analysis we will assume, that the transition matrix can be expressed as a function of a single continuous variable  $\theta$ . We are then computing the sensitivity  $\partial_\theta \mathbf{q}$  of the committor  $\mathbf{q}$  with respect to changes in the transition matrix  $\mathbf{T}(\theta)$  given in terms of the parameter  $\theta$ . It is shown, how this sensitivity can be used for a first-order estimate of the uncertainty in the committor  $\delta \mathbf{q}$  even in cases where the transition matrix  $\mathbf{T}$  is not exactly known, but is for example estimated from simulation data such as from molecular dynamics [54, 60].

#### 3.5.1 Sensitivity analysis

We first define the matrix

$$\mathbf{A} := \tilde{\mathbf{T}} - \text{Id},$$

so that the null space (kernel) of  $\mathbf{A}$  coincides with the space spanned by the two Perron eigenvectors of  $\tilde{\mathbf{T}}$ . This implies

$$\mathbf{A}\mathbf{q} = 0.$$

and the derivative with respect to  $\theta$  is

$$\partial_\theta \mathbf{A}\mathbf{q} = \mathbf{A}\partial_\theta \mathbf{q} + (\partial_\theta \mathbf{A})\mathbf{q} = 0. \quad (3.5.1)$$

To get  $\partial_\theta \mathbf{q}$  we need to invert  $\mathbf{A}$  which is not of full rank. Since  $\partial_\theta q_k = 0$  for  $k \in A \cup B$  by definition of the committor we can exclude these from the calculation and we then use the Moore-Penrose pseudoinverse  $[\cdot]^+$  (see appendix C) as a way to only invert the part of  $\mathbf{A}$ , that is not in the kernel and thus affecting states  $k \in A \cup B$ . This ensures the correct treatment of the boundary conditions and leads to the expression

$$\partial_\theta \mathbf{q} = -[\mathbf{A}]^+ (\partial_\theta \mathbf{A})\mathbf{q} \quad (3.5.2)$$

for the sensitivity.

### 3.5.2 Uncertainty / Sampling Error of the Committor

Let us now consider the case where the transition matrix  $\mathbf{T}$  is not known exactly but is instead sampled by a finite number of observations [54, 60, 64, 38]. We will be interested in the question of how the uncertainty involved in this finite sampling translates into uncertainty of the committor. We use the same definitions as in section 2.6 and let  $\mathbf{Z}$  be the count matrix of observed counts and  $\mathbf{Z}^{\text{Prior}}$  the matrix of virtual prior observations imposed on the model. We are looking for the sensitivity of the entries of the transition matrix  $T_{ij}$  to these count matrices.

The covariance between elements in different rows in the transition matrix is zero since a change in an element of the count matrix can affect only transition probabilities in the same row, provided that detailed balance is not enforced. This independence allows to treat the uncertainty for each row independently and we compute the sensitivity of  $\mathbf{A}$  to changes in the entries in  $\mathbf{T}$  by

$$\partial_j^{[i]} \mathbf{A} \equiv \partial_{T_{ij}} \mathbf{A} = \partial_{T_{ij}} \mathbf{T} = \mathbf{e}_i \cdot \mathbf{e}_j^T$$

leading to the expression for the sensitivity  $\mathbf{S}$  of the committor

$$\begin{aligned} \mathbf{S}_{[\cdot, j]}^{[i]} &\equiv \partial_j^{[i]} \mathbf{q} = -[\mathbf{A}]^+ \cdot (\mathbf{e}_i \mathbf{e}_j^T) \mathbf{q} \\ &= -[\mathbf{A}]_{[\cdot, i]}^+ q_j \end{aligned}$$

and in matrix form

$$\mathbf{S}^{[i]} = -[\mathbf{A}]_{[\cdot, i]}^+ \mathbf{q}^T.$$

The covariance between elements from the same row in the transition matrix  $\mathbf{T}$  is computed from the Dirichlet posterior distribution (see section 2.6) and the covariance matrix  $\Sigma^{[i]}$  becomes

$$\Sigma_{ab}^{[i]} \equiv \text{cov} [T_{ia} | T_{ib}] = \frac{(\alpha_a^{[i]})(\alpha^{[i]} \delta_{ab} - \alpha_b^{[i]})}{(\alpha^{[i]})^2((\alpha^{[i]} + 1))}$$

and where we used that

$$\alpha^{[i]} \equiv \mathbf{Z}_{[i,\cdot]}^{\text{eff}} + 1$$

and

$$\alpha^{[i]} \equiv \sum_{j=1}^M \alpha_j^{[i]} = z_i^{\text{eff}} + M$$

to simplify the notation. In matrix form this takes the form

$$\Sigma^{[i]} = ((\alpha^{[i]})^2(\alpha^{[i]} + 1))^{-1} \left( \text{diag} \left( \alpha^{[i]}(\alpha^{[i]}) \right) - (\alpha^{[i]})(\alpha^{[i]})^{\text{T}} \right) \quad (3.5.3)$$

and the standard deviation can be written as

$$\delta t_a = \sqrt{\text{cov} [T_{ia} | T_{ia}]} = \left( \frac{(\alpha_a^{[i]})(\alpha^{[i]} - \alpha_b^{[i]})}{(\alpha^{[i]})^2(\alpha^{[i]} + 1)} \right)^{1/2}.$$

Another simple and often used approach for propagating uncertainty from  $\mathbf{T}$  to the uncertainty of the committor (or any other property derived from  $\mathbf{T}$ ), is to sample the posterior distribution of transition matrices and compute the committor for each sample of this distribution [54, 64]. However, this procedure involves sampling itself and thus creates additional uncertainty in the estimation of the original uncertainty, which may be undesirable in situations where the estimation is conducted repeatedly, e.g. within an adaptive sampling scheme [54, 60].

To derive the final committor covariance we start with linear error propagation for the committor and use the sensitivity  $\mathbf{S}^{[i]}$ , given in Eq. (3.5.2), to extend the error in the transition matrix  $\Sigma^{[i]}$  as follows:

$$\begin{aligned} \text{cov} [\mathbf{q} | \mathbf{q}] &= \sum_{i \in \Omega} \mathbf{S}^{[i]} \cdot \Sigma^{[i]} \cdot (\mathbf{S}^{[i]})^{\text{T}} \\ &= \sum_{i \in \Omega} [\mathbf{A}]_{[i,\cdot]}^+ \left( \mathbf{q}^{\text{T}} \cdot \Sigma^{[i]} \cdot \mathbf{q} \right) \left( [\mathbf{A}]_{[i,\cdot]}^+ \right)^{\text{T}} \\ &= \sum_{i \in \Omega} \left( [\mathbf{A}]_{[i,\cdot]}^+ \cdot \left( [\mathbf{A}]_{[i,\cdot]}^+ \right)^{\text{T}} \right) \left( \mathbf{q}^{\text{T}} \cdot \Sigma^{[i]} \cdot \mathbf{q} \right) \end{aligned}$$

We then insert the analytical expression for the uncertainty  $\Sigma$  in Eq. (3.5.3) to obtain an expression for the covariance between single elements of the committor as

$$\begin{aligned}
\text{cov} [q_x | q_y] &= \sum_{i \in \Omega} (((\alpha^{[i]})^2 (\alpha^{[i]} + 1))^{-1} ([\mathbf{A}]_{xi}^+ [\mathbf{A}]_{yi}^+) \times \\
&\quad \times (\mathbf{q}^T (\text{diag} (\alpha^{[i]} (\alpha^{[i]})) - (\alpha^{[i]})(\alpha^{[i]})^T) \mathbf{q}) \\
&= \sum_{i \in \Omega} (((\alpha^{[i]})^2 (\alpha^{[i]} + 1))^{-1} ([\mathbf{A}]_{xi}^+ [\mathbf{A}]_{yi}^+) \times \\
&\quad \times (\mathbf{q}^T \text{diag} (\alpha^{[i]} (\alpha^{[i]})) \mathbf{q} - \mathbf{q}^T (\alpha^{[i]})(\alpha^{[i]})^T \mathbf{q}) \\
&= \sum_{i \in \Omega} (((\alpha^{[i]})^2 (\alpha^{[i]} + 1))^{-1} ([\mathbf{A}]_{xi}^+ [\mathbf{A}]_{yi}^+) \times \\
&\quad \times (\alpha^{[i]} \mathbf{q}^T \text{diag} (\alpha^{[i]}) \mathbf{q} - (\mathbf{q}^T \alpha^{[i]})^2)
\end{aligned}$$

This last expression is quadratic in the number of states leaving us with the inversion of  $\mathbf{A}$  as the most expensive operation of cubic order. Finally, we introduce abbreviations for

1. the contribution of each row to the statistical uncertainty by

$$u^{[i]} = (((\alpha^{[i]})^2 (\alpha^{[i]} + 1))^{-1} (\alpha^{[i]} \mathbf{q}^T \text{diag} (\alpha^{[i]}) \mathbf{q} - (\mathbf{q}^T \alpha^{[i]})^2),$$

2. and the plain sensitivity of the committor by

$$v_{xy}^{[i]} = ([\mathbf{A}]_{xi}^+ [\mathbf{A}]_{yj}^+)$$

and compute the variance in the elements of the committor  $q_x$  by

$$\begin{aligned}
\delta^2 q_x &\equiv \text{cov} [q_x | q_x] = \sum_{i \in \Omega} v_{xx}^{[i]} u^{[i]} \\
&= \sum_{i \in \Omega} (((\alpha^{[i]})^2 (\alpha^{[i]} + 1))^{-1} ([\mathbf{A}]_{xi}^+)^2 \times \\
&\quad \times (\alpha^{[i]} \mathbf{q}^T \text{diag} (\alpha^{[i]}) \mathbf{q} - (\mathbf{q}^T \alpha^{[i]})^2). \quad (3.5.4)
\end{aligned}$$

Lastly, the total uncertainty contribution vector  $\mathbf{w}$ , given by

$$w_i = \sum_{x \in \Omega} v_{xx}^{[i]} u^{[i]}, \quad (3.5.5)$$

contains the contribution of state  $i$  to the total uncertainty and can therefore be used in order to direct new simulations that are most promising in reducing this error [60].

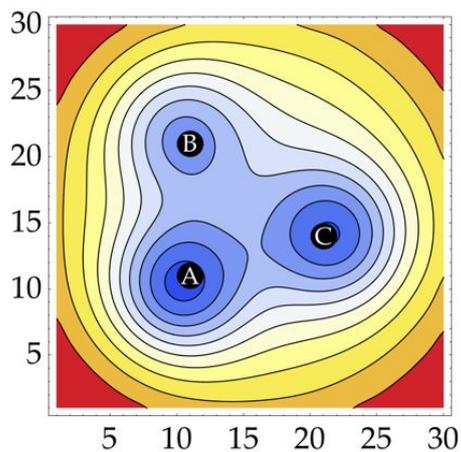


Figure 3.1 – Energy Landscape for diffusion in a 2D potential with 3 basins discretized into a grid of 30x30 bins. The minima in each basin are indicated by the letters *A*, *B* and *C*. Blue indicates low energies, red high energies.

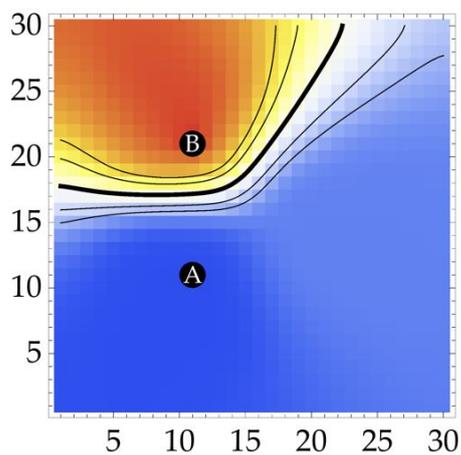


Figure 3.2 – 2D three-well model: Committor from state *A* to *B* computed directly from the reference transition probability matrix  $T_{ij}$ .

## 3.6 APPLICATIONS

## 3.6.1 Diffusion in a 2D Three-Well Potential

To illustrate an application of the above equations we use a simple model of a particle diffusing in a two-dimensional potential with three wells (see Figure 3.1 and appendix D for details), partitioned into a grid of  $M = 30 \cdot 30 = 900$ . The minima and their associated regions of configurational space will be referred to as  $A$ ,  $B$  and  $C$ . Transition probabilities are defined based on the potential energies  $U_i$  on each grid point using a Metropolis acceptance criterion given by

$$T_{ij} \propto \min \left( 1, \exp(-\beta(u_j - u_i)) \cdot \frac{\mathbb{P}(j \rightarrow i)}{\mathbb{P}(i \rightarrow j)} \right), i \neq j \quad (3.6.1)$$

with  $\beta = 1$ , which has the correct invariant distribution

$$\pi_i \propto \exp(-\beta U_i).$$

Only transitions between horizontal or vertical neighboring microstates are allowed with equal probability, resulting in a maximum of five nonzero entries per row in the 900x900 transition matrix. Note, that this choice requires to correct the Metropolis acceptance criterion for unsymmetric proposal probabilities  $\mathbb{P}(i \rightarrow j)$  in the edges and corners. This matrix is used as the reference for the dynamics of the system. The committor from state  $A$  to  $B$ , given in Eq. (3.3.9), is shown in Figure 3.2.

To investigate the dependence of the committor and its uncertainty on the actual number of observations and the chosen prior probability distribution, we computed the expected number of observed transitions in an equilibrium simulation as

$$\bar{Z}_{ij} = L \pi_i \hat{T}_{ij},$$

which is the product of the total number of simulation steps  $L$ , the invariant density of a state  $\pi_i$  and the true transition probabilities  $\hat{T}_{ij}$ . Four different types of prior distributions are considered here (see Tab. 3.1).

The committors computed for different simulation lengths

$$L = \{10^1, 10^3, 10^5, 10^7\}$$

and all prior sets except the null prior are presented in Figure 3.3.

The null prior was omitted since in this case the committor does not depend on the simulation length  $L$  and equals the exact committor (see Figure 3.2). It is important to note that this equivalence is only true on average and not for every possible simulation outcome. The influence of the full uniform prior is so strong that the computed

Prior	$Z^{\text{prior}}_{ij}$
Uniform Prior	0
Null Prior	-1
$1/M$ Prior	$1/M - 1$
Neighbor Prior	$\begin{cases} 0 & \text{if } (i,j) \text{ neighbors} \\ -1 & \text{else} \end{cases}$

Table 3.1 – Prior probability distributions used for the 2D example

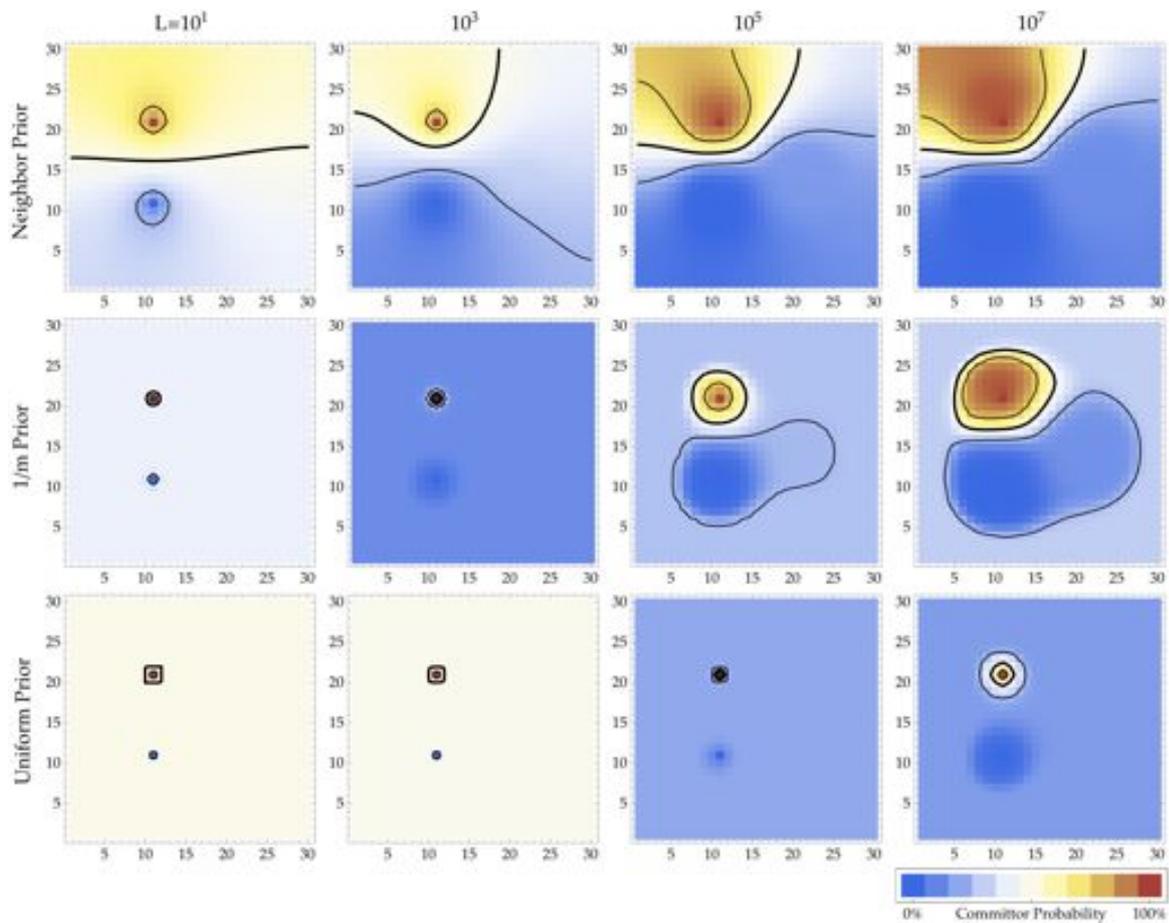


Figure 3.3 – 2D three-well model: Committor from state A to B computed for different prior choices (rows: neighbor prior,  $1/M$  prior, full uniform prior) and simulation lengths (columns:  $L = \{10^1, 10^3, 10^5, 10^7\}$ ). Isocommittor surfaces for  $q = \{0.25, 0.5, 0.75\}$  are given in black.

committor differs from the true committor vastly even for  $L = 10^7$ . The other two priors behave similarly to each other while the neighbor prior has the general advantage over the null prior that it always provides a transition matrix that can numerically be evaluated.

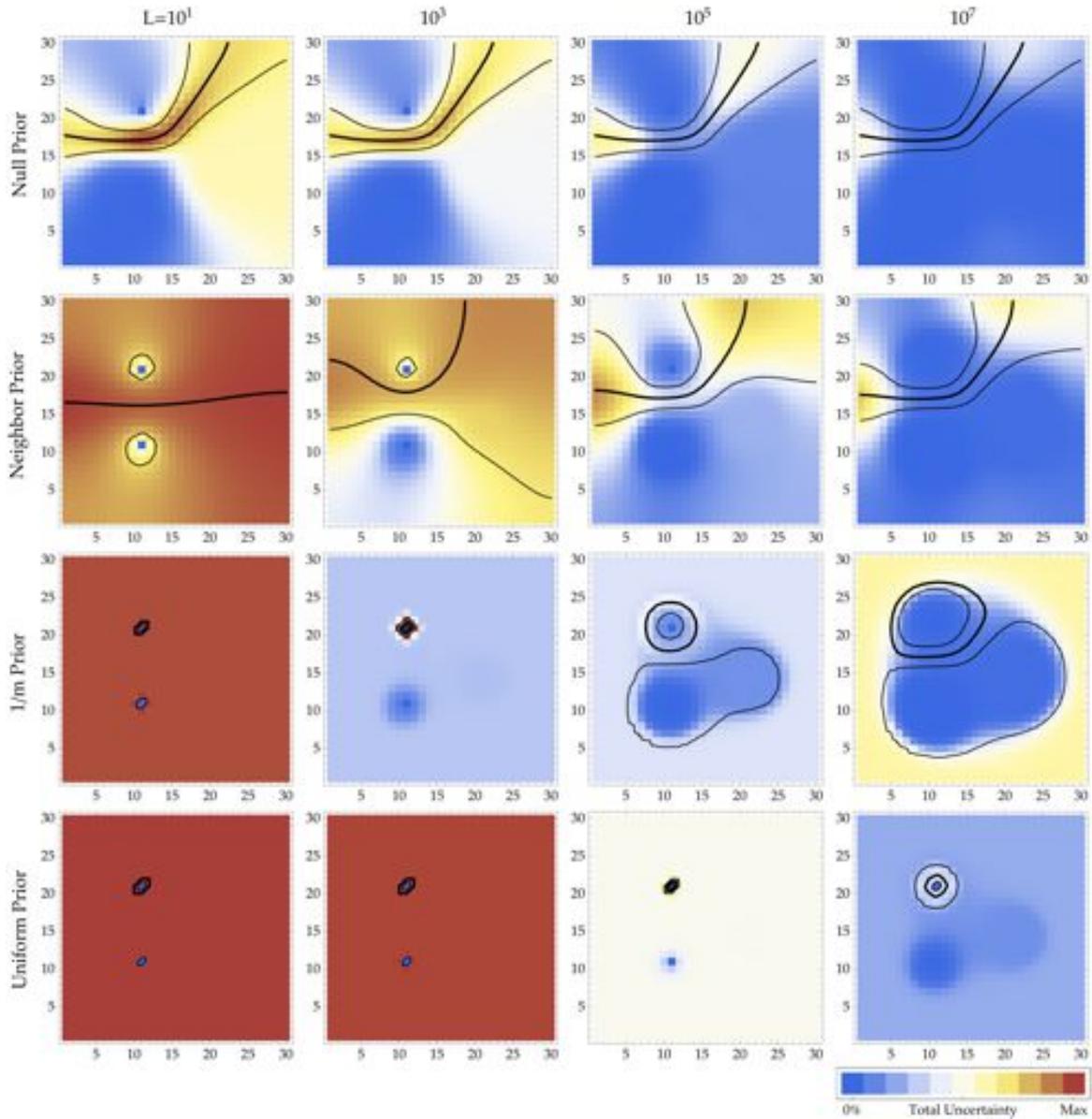
Eq. (3.5.4) gives the expression for the uncertainty in the computed average committor from a given number of observations. For the same set of total observations  $L$  and all priors in Table 3.1 the covariance was computed and is shown in Figure 3.4. The main uncertainty is always greatest in the transition region, and depends strongly on the choice of the prior, especially when few observations have been made.

Figure 3.5 show the difference in the predicted committors compared to the true reference committor given in Figure 3.2. The quality of the average predicted committor depends mainly on the amount of prior information put into the predictions: Priors with little information (null prior, neighbor prior) have less bias, while priors with much information ( $1/M$  prior, uniform prior) strongly bias the computed committor. However, committors with much information are less sensitive to perturbations in the transition matrix elements (see in Figure 3.6, i.e. thus having smaller uncertainties. However due to the bias this uncertainty is misleading in the cases of few observations. This behavior changes once the simulation length is long enough for the estimated committors to be similar.

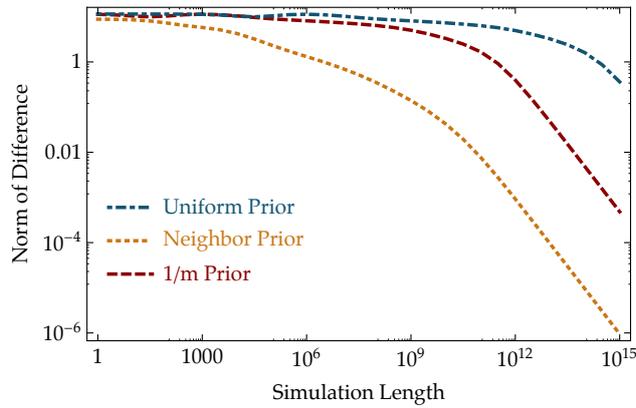
The effects of the bias of the prior are also visible in the contribution to the uncertainty from each state  $i$  given by  $w_i$  in Eq. (3.5.5) as shown in Figure 3.7. In general the main part of the contributions to the uncertainty is located in states inside the transition region. For small simulation lengths  $L$  the contribution is more widely distributed and mainly in regions that have also a significant equilibrium probability. With increasing simulation time, the uncertainty contributing states shift towards the outer perimeter of the energy landscape, where the uncertainty remains mostly unchanged since these parts of phase space are hardly visited at all.

The net flux for the system as given by Eq. (3.2.2) is shown in Figure 3.8. The opacity of the arrows indicates the intensity of the flux in the direction of the arrow. The main fraction of the flux traverses the barrier between A and B, while a minor fraction travels over state C.

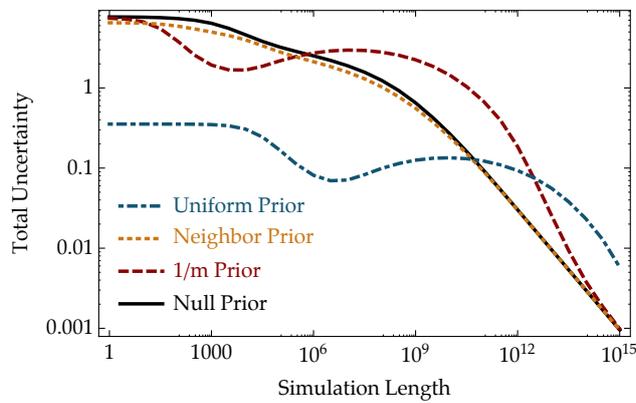
Finally, the 3-state committor, given by Eq. (3.4.1) was computed for states A, B and C (see Figure 3.9), thus partitioning the configurational space into three subsets divided by the main barriers. In this manner the multistate committor can be used to partition the configurational space into subsets, that are dynamically close to one state of a set of predefined states which can be regarded as cluster centers.



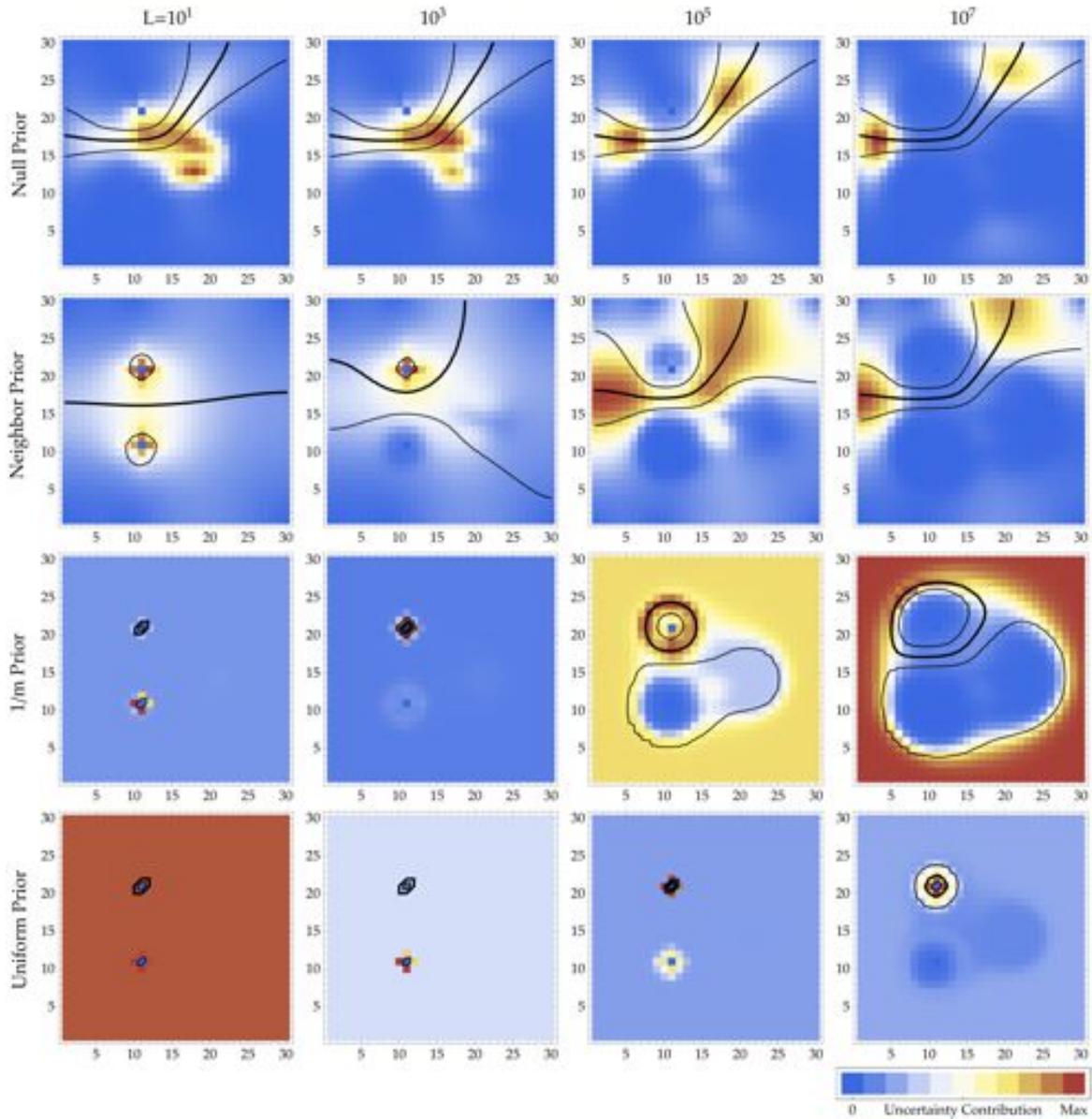
**Figure 3.4 – 2D three-well model: Statistical uncertainty (variance) in the entries of the committor probability  $\text{Cov}[q_x, q_x]$  in Eq. (3.5.4) from state A to B for different prior choices (rows: null prior, neighbor prior,  $1/M$  prior, full uniform) and simulation lengths (columns:  $L = \{10^1, 10^3, 10^5, 10^7\}$ ). Isocommittor surfaces from Figure 3.3 shown in black. Blue indicates no variance, red indicates high variance. The related absolute error development is given in Figure 3.6. States A and B are fixed by definition, thus at this points the variance is equal to zero. The highest variation is found in the transition region, the size of which depends strongly on the prior information. With increasing simulation length, the error in the low energy states reduces fastest.**



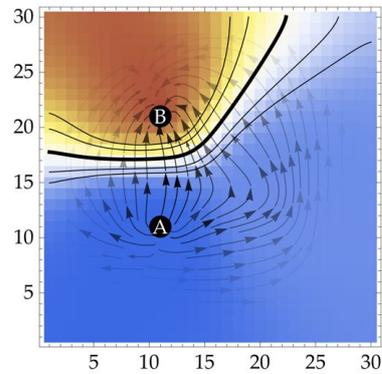
**Figure 3.5 – 2D three-well model: Norm of the difference of the computed committor** for different prior probability distributions (neighbor prior,  $1/M$  prior, full uniform) versus simulation length  $L$ . The uniform estimation is about six orders of magnitude slower in convergence since the amount of prior information is also about six orders of magnitude larger compared to the other priors.



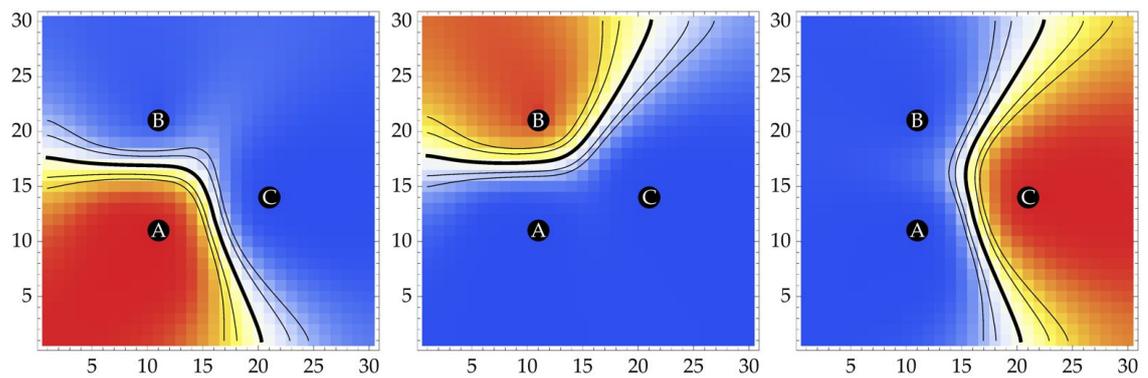
**Figure 3.6 – 2D three-well model: Theoretical average uncertainty in the estimated committor** for different prior probability distributions (null prior, neighbor prior,  $1/M$  prior, full uniform) versus simulation length  $L$ . The initial erratic behavior of the  $1/M$  prior and uniform prior is caused by a wrong committor prediction due to the high impact of these prior, when only few transitions have been observed.



**Figure 3.7 – 2D three-well model: Uncertainty contribution vector  $w_i$  in Eq. (3.5.5) for different prior choices** (rows: null prior, neighbor prior,  $1/M$  prior, full uniform) and simulation lengths (columns:  $L = \{10^1, 10^3, 10^5, 10^7\}$ ). Isocommittor surfaces from Figure 3.3 shown in black. Blue indicates vanishing sensitivity, red maximal sensitivity for each plot separately, thus absolute comparison is not possible between plots. This was chosen to better indicate the highest uncertainty contributions. The absolute sensitivity is given in Figure 3.6. The figure shows that in the case of the uniform prior a length of  $L = 10^7$  is insufficient for an accurate description of the uncertainty.



**Figure 3.8 – 2D three-well model: Net Flux between states A and B** computed from the reference transition matrix  $\hat{T}_{ij}$ . The underlying colors represent the reference committor. Arrows indicate the direction of the flux and the opacity the intensity. Most flux travels over the direct barrier from state A to state B.



**Figure 3.9 – Committor Computed for 3 states** from Eq. (3.4.1). The committor clearly shows a clear separation of the configurational space into 3 subsets divided by the potential barriers.

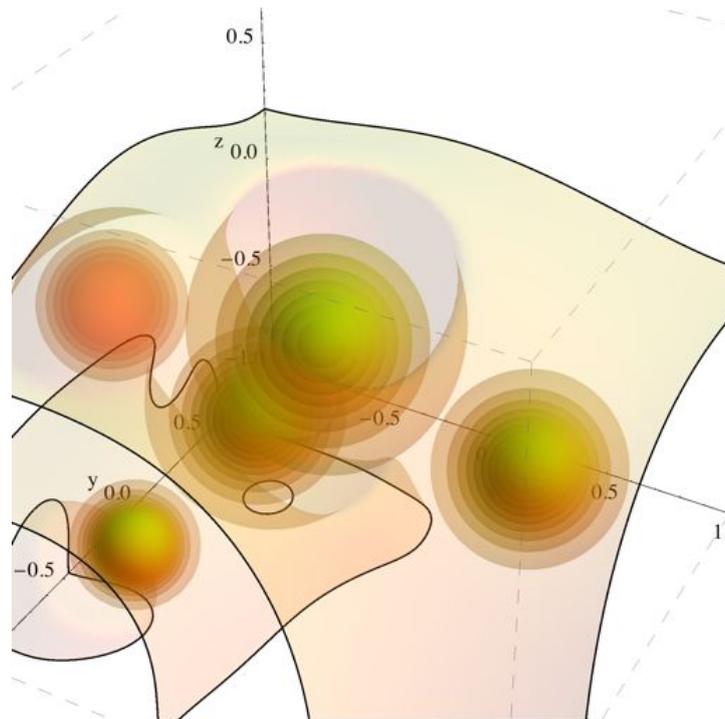


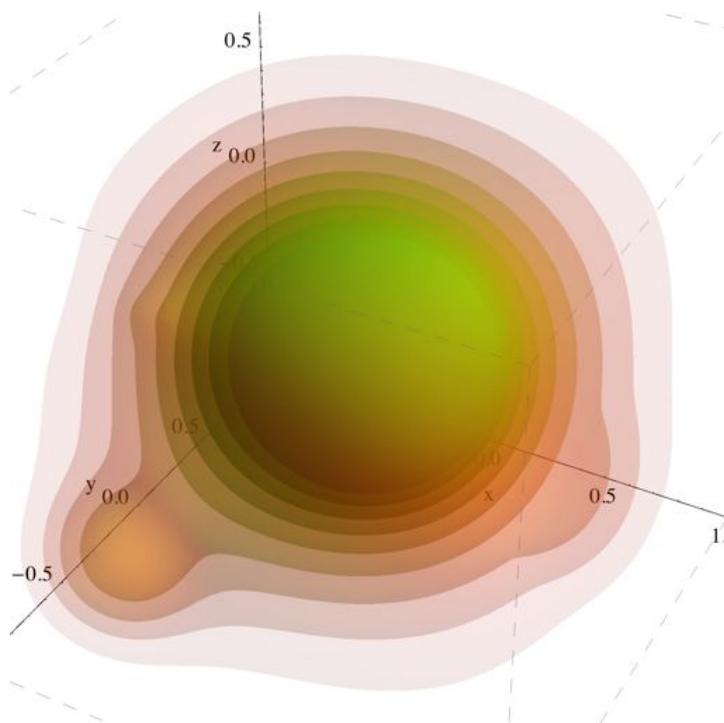
Figure 3.10 – 3D Ligand:protein model: Equipotential surfaces.

### 3.7 3D MODEL

The method is now further examined on a simple model system that mimics diffusional protein:ligand association. For this, a 3-dimensional potential was defined by a potential function  $U$  by a sum of five 3D-Gaussian functions that mimic an electrical field in which the ligand diffuses (for parameters see appendix D). The potential was coarse-grained on a grid with a total of  $M = 100 \cdot 100 \cdot 100 = 10^6$  states in the range of  $[-1, 1]^3$ , while the dynamics was modeled as a diffusional process under the influence of the potential as in the previous 2D case (see Eq. (3.6.1)). Figure (3.10) shows equipotential surfaces for a set of 19 exponentially spaced values of the potential  $U(x)$ , effectively depicting surfaces of equal equilibrium probability.

The outer boundary of the grid is defined as the “unbound” state A while all states inside a sphere at the center of the grid with a radius of 0.2 define the “bound” state B. The committor probability was computed using the procedure described in the theory section, employing the Power method to solve for the dominant eigenvector of the absorbing process [106]. The isocontours of the committor are shown in Figure 3.11. It is seen that these contours are roughly spherical around the binding site B, but have protrusions due to the existence of local energy minima.

Figure 3.12 shows some paths integrated along the normals to the isocommittor hypersurfaces. To compute these the committor function, given on each grid point, was interpolated by linear polynomi-



**Figure 3.11 – 3D Ligand:protein model: Isocommittor surfaces for the potential  $U$ .**

als between each neighboring grid point and computed the normals from the continuous interpolation. As initial points a set of 20 circularly positioned points on the inner B state were chosen which were directed toward the potential minimum at  $\{-0.6, -0.6, -0.6\}$  (see appendix D). The integrated paths define a bundle of field lines connecting the outer perimeter and the binding site, depicting the most probable paths towards the virtual binding site on the protein.

Using the committor also the reactivity  $g$  [63], *i.e.* the probability that a state contributes to a reactive trajectory, was computed using

$$g_i = q_i^+ \pi_i q_i^- \quad (3.7.1)$$

The results are shown in Figure 3.13. Due to the higher equilibrium probability in Eq. (3.7.1), the density of reactive trajectories increases towards the binding site and especially in the local minima.

### 3.8 DISCUSSION AND CONCLUSION

We have conducted a numerical study of the committor probability – the central mathematical object for characterizing dynamical processes – for discrete-state Markov processes.

An eigenvector-based approach to compute the committor probability was derived. This method is efficient and easy to implement and computes the committor for dynamical systems with large state

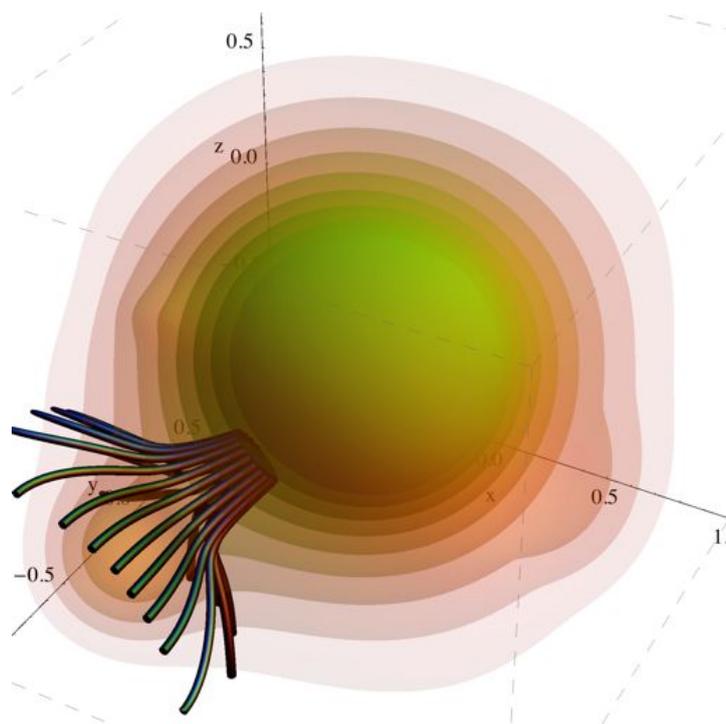


Figure 3.12 – 3D Ligand:protein model: Bundle of association path lines starting at the virtual binding site along the normals to the isocommittor surfaces.

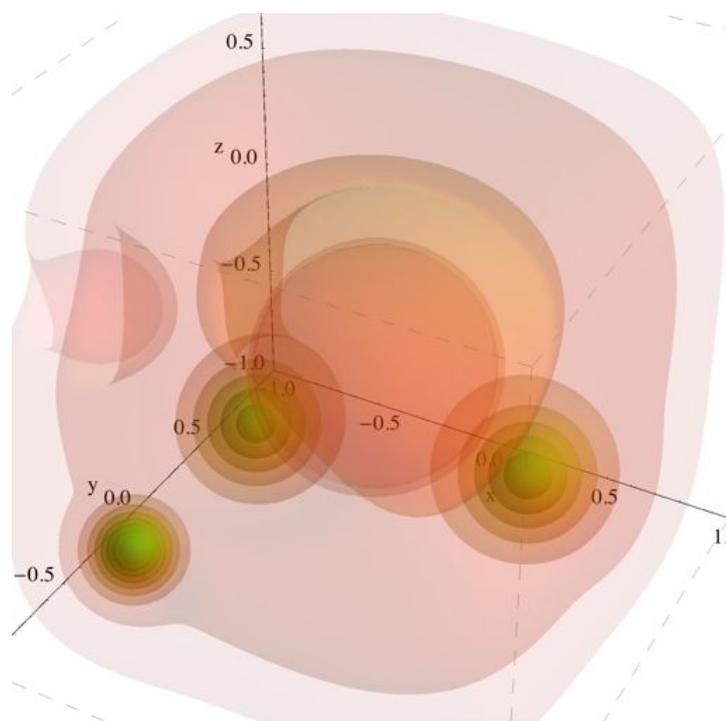


Figure 3.13 – 3D Ligand:protein model: Density of reactive trajectories  $g$  as given in Eq. (3.7.1).

spaces. If the considered transition matrix is sparse enough even very large systems can be investigated with computational effort roughly proportional to the number of states and is thus limited only by memory constraints. As an example, it was demonstrated that the approach is able to investigate the electrostatically steered ligand binding pathways to a protein which has also been used in [7].

Furthermore, a sensitivity and error analysis of the committor was conducted. Computation of the sensitivity requires the inversion of a matrix of the size of the number of states which is in general of cubic order, but can be made quadratic if the matrix is sufficiently sparse. The other computations are also maximally of quadratic order, which in principle allows a sensitivity analysis for medium system sizes.

Using the obtained error analysis, an adaptive algorithm can be defined for a fast committor computation by collecting information from different parts of the configurational space separately and combining this to produce more accurate estimations than possible from one single long simulation.



---

## MULTI-ENSEMBLE ESTIMATION

---

This chapter is based on the publications

- [3] Prinz, J.-H. Chodera, J. D., Pande, V. S., Swope, W. C., Smith, J. C. & Noé, F. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.* **134**, 244108 (2011). doi [10.1063/1.3592153](https://doi.org/10.1063/1.3592153).
- [4] Chodera, J. D., Swope, W. C., Noé, F., **Prinz, J.-H.**, Shirts, M. R. & Pande, V. S. Dynamical reweighting: improved estimates of dynamical properties from simulations at multiple temperatures. *J. Chem. Phys.* **134**, 244107 (2011). doi [10.1063/1.3592152](https://doi.org/10.1063/1.3592152).

### 4.1 INTRODUCTION

Sampling the underlying phase space by straightforward molecular dynamics simulation often suffers from the problem that the time scales of conformational changes can be orders of magnitude larger than simulation times accessible using current computational resources. Parallel Tempering (PT) molecular dynamics simulation has been an effective and thus popular approach to overcoming the issue of convergence in molecular simulations, by allowing replicas to heat up and overcome enthalpic barriers as the simulation proceeds while still sampling from an appropriate equilibrium distribution [110, 111, 112, 113]. At the same time this approach permits an analysis of the temperature dependence of properties of interest, which is especially important for comparisons with certain experimental results (e.g. melting curves, heat capacities) [57]. Although PT molecular dynamics produces unphysical replica trajectories, the short physical trajectories in between the exchanges can provide useful dynamical information. If the PT simulation is well-equilibrated, these initial configurations of the short trajectory segments will be sampled from the equilibrium at their corresponding temperatures.

Buchete and Hummer have shown that both thermodynamic and kinetic properties can be estimated over the range of temperatures by constructing Markov models using the short physical trajectories generated from PT simulations [114]. However, if a complete description

of dynamics across the entire thermodynamically relevant configurational space at a given temperature is desired, one quickly runs into problems if use is made of trajectories only from the temperature of interest, as some states that are sampled at other temperatures may not be well-sampled at the single temperature [114]. One would like to make use of the data collected at *all* temperatures to characterize the kinetic behavior in all regions sampled over the full range of temperatures spanned by the PT simulation in a manner similar to equilibrium reweighting [115, 116, 117, 118, 119, 120].

Here, we propose a method for integrating Molecular Dynamics (MD) simulation data from all temperatures by making use of *dynamical reweighting* [4], allowing a smooth, continuous, and differentiable estimate of the transition probabilities at any temperature without requiring the assumption of any kinetic model (such as Arrhenius kinetics [121]) and taking advantage of the increased transition rates at higher (or, for transitions with entropic barriers, lower) temperatures. Reweighting methods (such as histogram-based [115, 116, 117, 118] or histogram-free [119, 120] approaches) allow the use of samples collected from multiple distributions to provide an improved estimate of the expectation value of some static property at the distribution of interest, and have been used extensively in the analysis of equilibrium thermodynamic properties in replica-exchange simulations [122].

Dynamical reweighting has recently been proposed as a new way of estimating dynamical properties (such as correlation functions) using an asymptotically optimal estimator, simultaneously providing a good estimate of the statistical error [4]. Here, we show how dynamical reweighting can be used to estimate transition probabilities (and their statistical uncertainties) for the construction of a Markov model as a smooth function of temperature, making use of data from all temperatures. This has the advantage of producing a useful Markov model at *any* temperature containing the dependence of kinetic properties on temperature while providing an assessment of the error in this model.

We illustrate this approach for the standard test case of the terminally-blocked alanine peptide in explicit solvent. A Markov model constructed from short (6 ps) trajectories from each state has been previously shown to accurately describe the kinetics of this system at 302 K [65]. This peptide system presents a challenge for estimators based on individual temperatures, due to the presence of highly metastable states with very high free energies relative to the most populated states. These states are poorly sampled at temperatures near 300 K, even though their temporal behavior can dominate the non-equilibrium relaxation kinetics at this temperature. Finally, we also determine whether dynamical reweighting using all the data produces substantially improved kinetic models in comparison to (a) present single-temperature approaches, such as Buchete and Hum-

mer [114] in the full range of temperature and (b) to a high-statistics model at 302 K.

## 4.2 DYNAMICAL REWEIGHTING

The recently proposed method named *dynamical reweighting* [4] allows to compute correlation function estimates of a system where observed time series have been collected from different ensemble settings. In this section we lay out the necessary theoretical basis to compute these enhanced estimates for the specific case of a system in the canonical ensemble at several inverse temperatures  $\beta_k \equiv (k_B T_k)^{-1}$ , where the system obeys Hamiltonian dynamics meaning that any initial point

$$q(0) \equiv (q(0) \in \Omega, v(0))$$

in phase space is transported deterministically along a trajectory  $q(t)$  to a final point

$$q(\tau) \equiv (x(\tau) \in \Omega, v(\tau)).$$

This corresponds to the case of “canonical distribution of Hamiltonian trajectories” examined in detail in Ref. [4].

Within this system at a fixed temperature we assume that the Hamiltonian dynamics can be well approximated by a Markov model (as laid out in chapter 2). We now demonstrate how transitions observed at all temperatures can be used to infer transition probabilities at any temperature of interest through the use of a reweighting procedure.

With this definition of dynamics, the (now temperature-dependent) state-to-state correlation functions  $\mathbf{C}(\tau; \beta)$  for the crisp membership definition

$$\chi_i(x) : i \in \mathbf{M}, x \in \Omega \mapsto \{0, 1\}$$

can be expressed as a Boltzmann-weighted expectation function

$$C_{ij}(\tau; \beta) = \frac{1}{Z(\beta)} \int d q(0) e^{-\beta \mathcal{H}(q(0))} \chi_i(x(0)) \chi_j(x(\tau)),$$

where  $Z(\beta)$  is the complete partition function of both kinetic and potential energies. From these correlation functions we can compute the associated transition probabilities

$$T_{ij} = \frac{C_{ij}(\tau; \beta)}{\sum_{k \in \mathbf{M}} C_{ik}(\tau; \beta)}$$

as shown in Eq. (2.5.2).

Suppose we have a set of  $N$  Hamiltonian phase space trajectories

$$q^{[n]}(t), n \in \{1, \dots, N\}$$

of length  $t_{\max}$ ,  $t \in [0, t_{\max}]$  sampled from equilibrium at  $K$  distinct temperatures  $\beta_k$ ,  $k \in \{1, \dots, K\}$ . For convenience, we group the  $N$

trajectories into subsets  $Q_k \subset \{1, \dots, N\}$  according to which temperature  $\beta_k$  their initial phase space points  $q^{[n]}(0)$  were drawn from. For the reweighting procedure, only the total number of trajectories  $N_k \equiv |T_k|$  sampled from the temperature  $\beta_k$  is relevant, and not the direct association of a specific trajectory  $q^{[n]}(t)$  with the temperature  $\beta_k$  it was sampled from [4, 120].

By the application of *dynamical reweighting* [4], a correlation function  $C_{ij}(\tau; \beta)$  can then be estimated using the entire set of trajectories at all temperatures with

$$\hat{C}_{ij}(\tau; \beta) \approx \sum_{n=1}^N w^{[n]}(\beta) \hat{C}_{ij}^{[n]}(\tau; \beta) \quad (4.2.1)$$

where the individual trajectory contributions

$$\hat{C}_{ij}^{[n]}(\tau; \beta) = \text{cor} [\chi_i | \chi_j] (\tau)$$

are estimated as the state-to-state correlation functions as in Eq. (2.5.3). The normalized trajectory weights  $w^{[n]}(\beta)$  are given by

$$w^{[n]}(\beta) = \hat{Z}^{-1}(\beta) \left( \sum_{k=1}^K N_k \hat{Z}_k^{-1} e^{-(\beta_k - \beta) E_n} \right)^{-1}, \quad (4.2.2)$$

where the normalization constants  $\hat{Z}(\beta)$  are chosen that

$$\sum_{n=1}^N w^{[n]}(\beta) = 1$$

holds. Here,  $E^{[n]} \equiv \mathcal{H}(q^{[n]}(0))$  denotes the total energy of the system in trajectory  $q^{[n]}(t)$ , which is constant over trajectories subject to Hamiltonian dynamics. The missing normalization constants  $\hat{Z}_k$  need to be determined from the solution of a set of  $K$  self-consistent equations indexed by  $i \in \{1, \dots, K\}$

$$\hat{Z}_i = \sum_{n=1}^N \left( \sum_{k=1}^K N_k \hat{Z}_k^{-1} e^{-(\beta_k - \beta_i) E_n} \right)^{-1} \quad (4.2.3)$$

which can be obtained efficiently in a number of ways (see details in appendix B.1), although it is often necessary to work with logarithmic representations to avoid numerical instability. The choice of weights  $w^{[n]}$  in Eq. (4.2.2) gives an asymptotically optimal (i.e. lowest variance) estimate of the correlation function in Eq. (4.2.1). A detailed exposition of dynamical reweighting for the estimation of correlation functions is presented in Ref. [4].

Finally, the row-stochastic transition matrix estimate  $\hat{\mathbf{T}}(\tau; \beta)$  is computed from Eq. (4.2.1) using Eq. (2.5.3), where the symmetry of  $\hat{\mathbf{C}}(\tau; \beta)$  results in a reversible transition matrix estimate  $\hat{\mathbf{T}}(\tau; \beta)$  (i.e., it will satisfy detailed balance). Note, that this will introduce the unavoidable

projection error discussed in section 2.4.3. As stated above, we will assume that the lagtime  $\tau$  is large enough for this error to be small. If we would only be interested in certain properties there exist ways to circumvent this projection error as shown in chapter 5.

#### 4.2.1 Estimation of uncertainties in transition probabilities

For a given temperature  $\beta$  and lagtime  $\tau$ , the statistical uncertainty in  $\hat{C}_{ab} \equiv \hat{C}_{ab}(\tau; \beta)$  can be estimated in a straightforward manner [4, 120]. We start with the weight matrix

$$\mathbf{W} : n \in \{1, \dots, N\}, k \in \{1, \dots, K\} \mapsto W_{nk} \equiv w_n(\beta_k)$$

where  $k$  runs over the set of temperatures  $\beta_k$  and  $n$  over all trajectories. Augmenting  $\mathbf{W}$  by three additional columns, indexed by the letters A, X, and Y, defined by

$$\begin{aligned} W_{nA} &= w_n(\beta) \\ W_{nX} &= W_{nA} \frac{\hat{C}_{ab}^{[n]}}{\hat{C}_{ab}} ; & W_{nY} &= W_{nA} \frac{\hat{C}_{a'b'}^{[n]}}{\hat{C}_{a'b'}} \end{aligned}$$

the covariance  $\Sigma$  of two correlation matrix elements  $\hat{C}_{ab}$  and  $\hat{C}_{a'b'}$  can be estimated by [4]

$$\begin{aligned} \Sigma_{[a,b,a',b']} &\equiv \text{cor} [\hat{C}_{ab} | \hat{C}_{a'b'}] \\ &\approx \hat{C}_{ab} \hat{C}_{a'b'} \cdot (\hat{\Theta}_{AA} - \hat{\Theta}_{AY} - \hat{\Theta}_{XA} + \hat{\Theta}_{XY}). \end{aligned} \quad (4.2.4)$$

The covariance matrix estimate  $\hat{\Theta}$  is computed as [4]

$$\hat{\Theta} \equiv \mathbf{W}^T [\text{Id} - \mathbf{W}\mathbf{N}\mathbf{W}^T]^+ \mathbf{W}, \quad (4.2.5)$$

with  $\mathbf{N} = \text{diag}(N_1, \dots, N_K, 0, 0, 0)$  and  $[\cdot]^+$  the generalized inverse.

Next, the uncertainty  $\text{Var}(\hat{\mathbf{T}})$  in the estimated transition probabilities can be approximated as a function of the uncertainty in the correlation matrix elements,  $\Sigma_{[a,b,a',b']}$ , by a first-order Taylor expansion about the mean  $\langle \hat{T}_{ij} \rangle$ ,

$$\begin{aligned} \text{Var}(\hat{\mathbf{T}}) &\equiv \langle (\hat{\mathbf{T}} - \langle \hat{\mathbf{T}} \rangle)^2 \rangle \\ &\approx \sum_{a,a',b,b' \in \mathbf{M}} \left( \partial_{\hat{C}_{ab}} \hat{\mathbf{T}} \right) \Sigma_{[a,b,a',b']} \left( \partial_{\hat{C}_{a'b'}} \hat{\mathbf{T}} \right). \end{aligned} \quad (4.2.6)$$

Using Eq. (2.5.3), the sensitivity of  $\hat{\mathbf{T}}$  to the correlation matrix element  $\hat{C}_{ab}$  is given by

$$\frac{\partial \hat{T}_{ij}}{\partial \hat{C}_{ab}} = \frac{\delta_{aj} \delta_{bi} + \delta_{ai} \delta_{bj} - \delta_{ab}}{\sum_{k \in \mathbf{M}} \hat{C}_{ik}} - \frac{\hat{C}_{ij} (-M \delta_{ab} + \delta_{ai} + \delta_{bi})}{(\sum_{k \in \mathbf{M}} \hat{C}_{ik})^2}, \quad (4.2.7)$$

or in matrix form

$$\begin{aligned} \partial_{\hat{C}_{ab}} \hat{\mathbf{T}} &= \text{diag}(\hat{C}_i)^{-1} (\mathbf{e}_a \mathbf{e}_b^T + \mathbf{e}_b \mathbf{e}_a^T - \text{Id} \delta_{ab}) \\ &\quad - \text{diag}(\hat{C}_i)^{-2} (\hat{C} (-M \delta_{ab} + \mathbf{1} (\mathbf{e}_a^T + \mathbf{e}_b^T))). \end{aligned} \quad (4.2.8)$$

The complete expression for the variance in the transition probabilities  $\delta^2 \hat{T}_{ij}$  in Eq. (4.2.6) can be evaluated using Eqs. (4.2.4), (4.2.5) and (4.2.7) and the final result is complex but still calculable, as demonstrated in section 4.4. A detailed description of the procedure for computing statistical uncertainties for arbitrary properties can be found in Ref. [4, 120].

#### 4.2.2 Modified parallel tempering protocol

We employ a modified PT protocol in which a set of  $K \times N$  Hamiltonian trajectory segments  $q^{[k,n]}(t)$  of uniform length  $t_{\max}$  is generated [4] and with temperature index  $k \in \{1, \dots, K\}$  and iteration  $n \in \{1, \dots, N\}$ . For the Markov property to hold in the later estimation, the length  $t_{\max}$  between proposed exchanges in the PT protocol needs to exceed the lagtime  $t_{\max} \geq \tau$ . We start by assuming that some process was used to generate the initial phase space points  $q^{[k,1]}(0)$  for the first set of trajectories ( $n = 1$ ) from equilibrium within the canonical (NVT) ensemble at each corresponding inverse temperature  $\beta_k$

$$\mathbb{P} \left( q^{[k,1]}(0); \beta_k \right) = \frac{1}{Z(\beta_k)} \exp(-\beta_k \mathcal{H}(q^{[k,n]}(0))).$$

These initial phase space points may be obtained, for example, by a standard PT protocol, or by running the modified protocol for a number of iterations starting from one or more arbitrary initial configurations.

In each subsequent iteration ( $n > 1$ ), Hamilton's equations of motion are used to propagate all replicas using a symplectic integrator with sufficiently small timesteps to generate trajectories  $q^{[k,n]}(t)$  of length  $t_{\max}$ . At  $t = t_{\max}$ , we propose exchanges between the final configurations  $q^{[i,n]}(t_{\max})$  and  $q^{[i \pm 1, n]}(t_{\max})$  of neighboring temperatures  $\beta_i$  and  $\beta_{i \pm 1}$ , starting from the highest temperature down to the lowest one in odd iterations and in reverse order in even ones [65].<sup>1</sup> The Metropolis-like probability [113] for accepting or rejecting the exchange  $p_{\text{exch}}(i \leftrightarrow j)$  depends on the potential energies of the final configurations  $U_k \equiv U(Z_{kn}(t_{\max}))$  with

$$p_{\text{exch}}(i \leftrightarrow j) = \min \{1, \exp(-(\beta_i - \beta_j)(U_j - U_i))\}.$$

<sup>1</sup> Note that other exchange proposal schemes can be used, provided the resulting algorithm satisfies the condition of "balance" (which is less strict than detailed balance) [123].

Regardless of whether the exchange is accepted or rejected, we reassign the velocities according to the Maxwell-Boltzmann distribution [67] at the new (or old, if rejected) temperatures, and denote the new phase space points from which the next iteration can be carried out as  $q^{[k,n+1]}(0)$  (see proof in Appendix B.2). This satisfies the conditions defined by Okamoto [113] in order for the kinetic energies to not appear in  $p_{\text{exch}}(i \leftrightarrow j)$  and is equivalent to rescaling the velocities for accepted exchanges and then applying a massive collision for the Andersen thermostat [67]. The reason for reassignment of velocities instead of rescaling is that when using Hamiltonian trajectories to propagate dynamics, no thermostating would otherwise take place. Without this velocity reassignment, the use of Hamiltonian trajectories (even if velocity rescaling were performed after exchanges) would mean that no or minimal thermostating would take place, generating an improper ensemble; velocity reassignment ensures the canonical ensemble is generated.

### 4.3 APPLICATION TO MARKOV MODELS

In the previous section we have presented the statistical justification and the mathematical basis to use dynamical reweighting for the estimation of transition probabilities between a set of  $M$  states defined by  $\chi$ .

For reasons of comparability, we also consider two Bayesian methods for estimation of the transition matrices and rate matrices using data collected only from a single temperature. Both methods sample transition probabilities or rates according to the same likelihood function in Eq. (2.6.3) assuming detailed balance  $\pi_i T_{ij} = \pi_j T_{ji}$ , but employ different model parameterizations and, more importantly, different prior probability distributions.

#### 4.3.1 Reversible Transition Matrices

We first consider the approach described in Ref. [64] to infer transition matrices that satisfy detailed balance. Starting with an observation represented by the fractional count matrix  $\mathbf{Z}$ , the posterior probability a transition matrix  $\mathbf{T}$  was responsible for generating this observation is given by

$$\mathbb{P}(\mathbf{T}|\mathbf{Z}^{\text{eff}}) \propto \mathbb{P}(\mathbf{Z}|\mathbf{T}) \mathbb{P}(\mathbf{T}) = \prod_{i,j \in \mathcal{M}} T_{ij}^{Z_{ij}} \mathbb{P}(\mathbf{T}). \quad (4.3.1)$$

As the prior,  $\mathbb{P}(\mathbf{T})$ , we choose the “null”-prior (see section 2.6.4)

$$\mathbb{P}(\mathbf{T}) \equiv \prod_{i,j \in \mathcal{M}} T_{ij}^{-1} \quad (4.3.2)$$

Method	Abbreviation	Color
Transition matrix estimation [64]	[TE]	red
Rate matrix estimation [114]	[RE]	green
Dynamical reweighting [4]	[DR]	blue
Shooting trajectories at 302 K [65]	[ST]	black

**Table 4.1 – Estimation Methods** Table of methods used for transition probability or rate estimation with their corresponding abbreviations and colors used consistently throughout this paper.

to keep the influence from the prior as small as possible. Here, the distribution in Eq. (4.3.1) was sampled using a Markov Chain Monte Carlo (MCMC) procedure described in Ref. [64].

#### 4.3.2 Reversible Rate Matrices

We also consider a second approach to estimating transition probabilities from individual temperatures using Bayesian estimation. Here, we sample rate matrices  $\mathbf{K}$  (discussed in section 2.3) with elements  $K_{ij} > 0$  for  $i \neq j$  and  $K_{ii} = -\sum_{j \neq i} K_{ij}$  using the approach described in Ref. [114]. This approach does not estimate the transition probabilities directly, but instead uses a parametric form of a reversible rate matrix  $\mathbf{K}$ , that uses the logarithms of the elements in the upper-right triangular matrix  $K_{ij}$  for  $j > i$  (without diagonal entries) and the equilibrium distribution  $\pi_i$ , thus assuring a rate matrix with non-positive eigenvalues and also positive off-diagonal rates.

The posterior in Eq. (4.3.1), written in terms of the rate matrix  $\mathbf{K}$ , is given by

$$\mathbb{P}(\mathbf{K}|\mathbf{Z}^{\text{eff}}) \propto \mathbb{P}(\mathbf{Z}|\mathbf{K})\mathbb{P}(\mathbf{K}) = \prod_{i,j \in \mathbf{M}} \exp(\tau \mathbf{K})_{ij}^{Z_{ij}} \mathbb{P}(\mathbf{K}) \quad (4.3.3)$$

where the prior is uniform in  $\ln K_{ij}$ ,  $j > i$  and  $\ln \pi_i$ . After sampling rate matrices with a Metropolis Monte Carlo scheme [114] the related set of transition matrices with the lagtime  $\tau$  is computed by

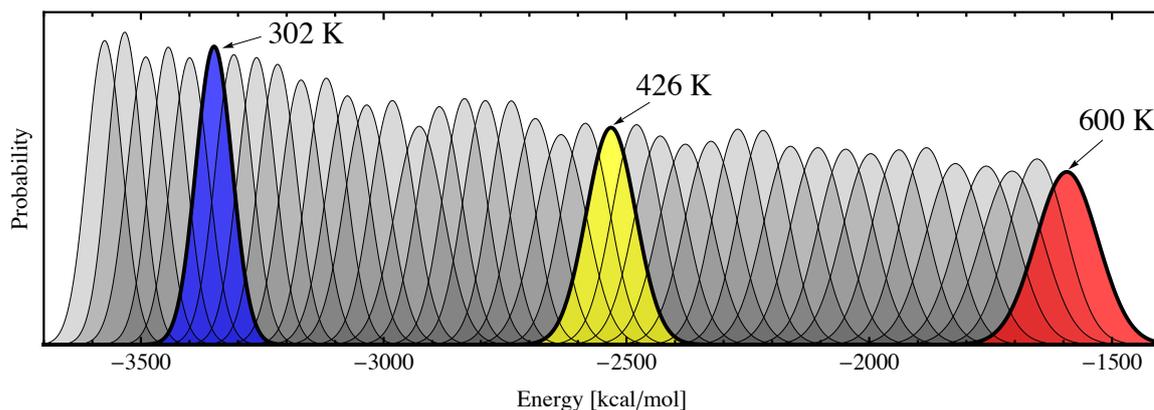
$$\mathbf{T}(\tau) = \exp(\tau \mathbf{K}) \quad (4.3.4)$$

All methods with their abbreviations and colors used consistently in this chapter are listed in Table 4.1.

## 4.4 APPLICATION TO ALANINE DIPEPTIDE PT SIMULATION

### 4.4.1 System Setup

To illustrate the construction of temperature-dependent Markov models using dynamical reweighting, we estimated the transition proba-



**Figure 4.1** – Distribution of trajectory total energies from parallel tempering simulation. Distributions for trajectories sampled at temperatures 302 K, 425 K, and 600 K — for which single-temperature Bayesian analysis convergence properties are shown in Figure B.1 — are highlighted.

bilities between conformational states for the terminally-blocked alanine peptide (Ace-Ala-Nme) (see Figure 4.2) in explicit solvent from a PT molecular dynamics simulation. This dataset<sup>2</sup> was published previously as part of a study that explored the suitability of a Markov model for describing dynamics at 302 K [65]; here, we make use of this dataset to facilitate comparison to an independent trajectory set that appears in that reference. The PT dataset consists of an ensemble of 501 Hamiltonian trajectories 20 ps in length at each of 40 temperatures, spanning 273 to 600 K, with peptide configurations stored every 0.1 ps. The temperatures were exponentially spaced to ensure good overlap in the potential and total energy distributions between neighboring temperatures and reasonably high exchange probabilities (see Figure 4.1).

A leapfrog Verlet integrator [124, 125, 126] (with bonds involving hydrogen atoms constrained) was used to produce the dynamical trajectories. The fluctuation in total energy averaged over all 20 ps trajectories at each temperature was minimal and the drift negligible (see Table 4.2). The production run followed a 1 ns equilibration phase during which exchanges were attempted at 1 ps intervals, ensuring that all initial configurations were drawn from equilibrium at their respective temperatures.

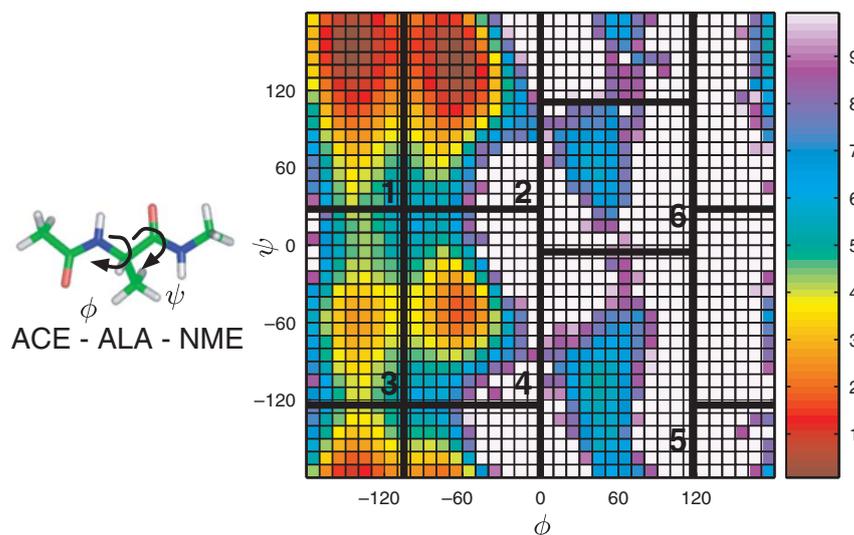
Previous work has demonstrated that a Markov model based on a six-state decomposition ( $M = 6$ ), as depicted in Figure 4.2, can accurately describe the dynamics of this peptide for lagtimes longer than  $\tau = 6$  ps [65]. We employ the same state decomposition for all temperatures with the suggested minimal lagtime of  $\tau = 6$  ps.

To evaluate the accuracy of the methods for estimating transition probabilities, we compare the separate estimates obtained using Dynamical reweighting (DR), Transition Matrix Estimation (TE), and Rate

<sup>2</sup> The alanine dipeptide parallel tempering and kinetics datasets are available online for download at <https://simtk.org/home/alanine-dipeptide/>.

Temp [K]	SD [kcal/mol]	Drift [kcal/(ps mol)]
302	$0.214 \pm 0.014$	$0.0056 \pm 0.0006$
426	$0.280 \pm 0.019$	$0.0073 \pm 0.0006$
600	$0.376 \pm 0.026$	$0.0097 \pm 0.0011$

**Table 4.2 – Consistency of (PT) trajectories** Standard deviation (SD) and drift of the total energy over 20 ps leapfrog trajectories, averaged over all trajectories at selected temperatures.



**Figure 4.2 – Terminally-blocked alanine peptide and potential of mean force with Markov state definitions.** Left: The terminally blocked alanine peptide with  $(\phi, \psi)$  torsions labeled. Right: The potential of mean force as a function of  $(\phi, \psi)$  torsions at 302 K in units of  $k_B T$ , estimated from the PT simulation using WHAM [83, 118], truncated at  $10 k_B T$  (white regions). The six manually identified states are labeled in black (note periodic boundaries). The picture was taken from Ref. [65].

Matrix Estimation (RE) with a separate dataset of  $6 \times 10\,000$  short (10 ps) trajectories (ST) initiated from the equilibrium ensemble within each state at 302 K, also taken from Ref. [65]. The PT simulation, in comparison, furnishes a total of 501 independent trajectories at that temperature.

The system is small enough that reasonable statistics can be obtained with moderate CPU requirements, while complex enough that some transitions (and even some states) are sampled only at high temperatures. In what follows, the results from the Markov model obtained from the dynamical reweighting method are compared to the model computed by Bayesian analyses using data from *individual* temperatures from the parallel tempering simulation, as in Buchete and Hummer [114].

#### 4.4.2 Estimated transition probabilities as a function of temperature

A comparison of various approaches to estimating the transition probabilities between all  $6 \times 6$  pairs of states as a function of temperature is given in Figure 4.3. The blue lines give the estimates from dynamical reweighting (DR) [65] using all available data at all temperatures, as described in Section 4.2. To obtain the normalization constants  $\hat{Z}_k$ , we solved the set of self-consistent equations in Eq. (4.2.3) with a relative convergence tolerance in the residual of  $10^{-7}$  (see Appendix B.1). Transition probabilities were also estimated at one intermediate temperature between each pair of simulated temperatures.

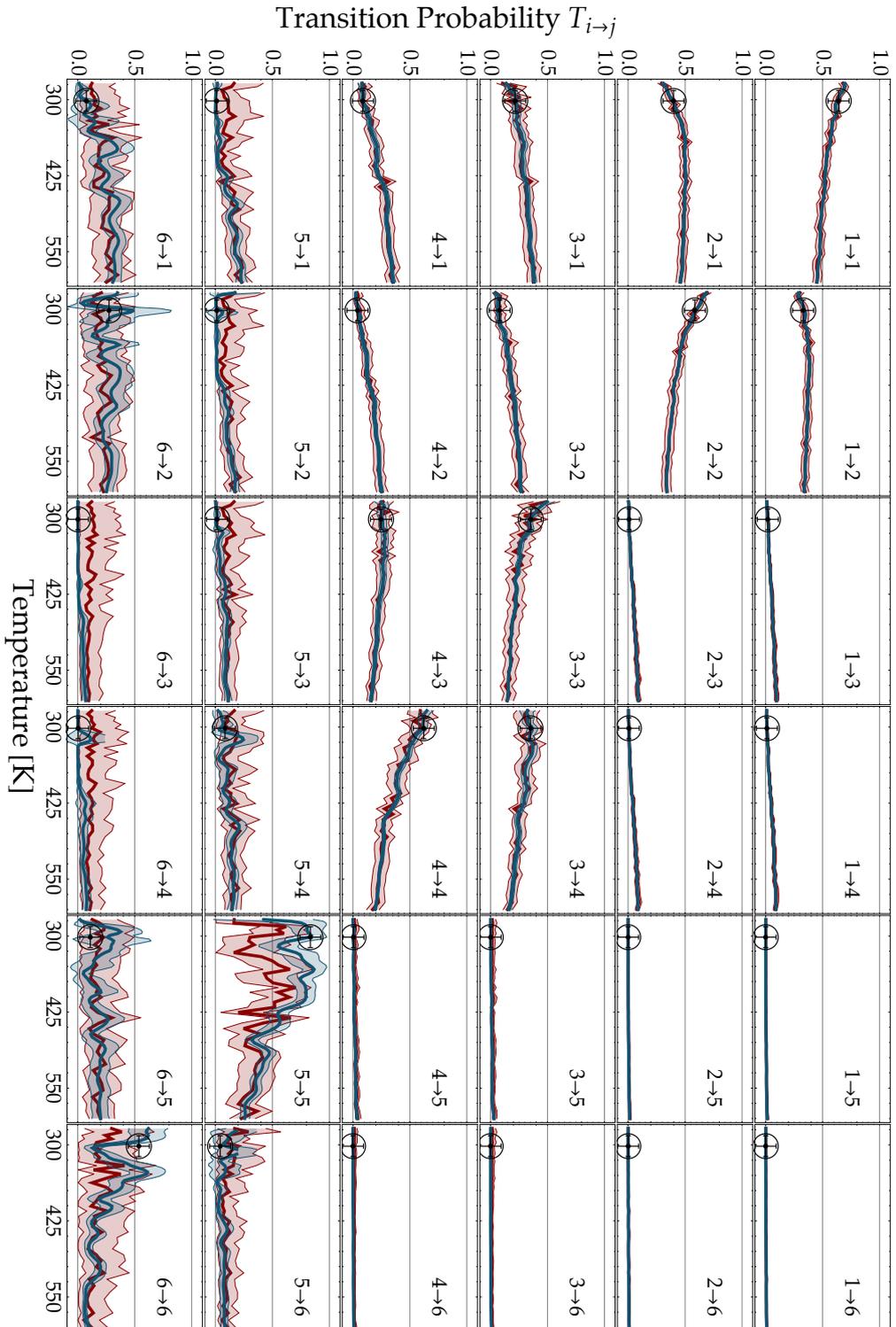
The red lines in Figure 4.3 show transition probabilities for the reversible single temperature estimation of transition matrices (TE) [64]. For each of the 40 temperatures the sampler was run to collect a total of 10 000 samples. For the sampling of reversible rate matrices (RE) the same amount of data was collected using the sampling proposed in Ref. [114]. Diagnostics of convergence for both methods appear as supplementary Figure B.1. For reasons of clarity, the performance of the rate matrix estimation (RE) is only shown in the detailed comparison plots in Figure 4.4 discussed in the subsequent section. The black cross-hair in Figure 4.3 refers to the reference values (ST) at 302 K estimated from the shooting trajectories.

Qualitatively, both methods agree, especially for transitions among highly populated states (1,2,3,4). However, the dynamical reweighting estimate, which uses the combined data from all temperatures, has smaller uncertainties than the estimators that use only individual temperatures. In addition, the general agreement with the reference simulation is best for dynamical reweighting.

#### 4.4.3 Detailed comparison of transition probability estimates at 302 K

For a detailed comparison with precisely known transition probabilities, the Bayesian analysis method with reversibility constraint for transition matrices [64] (TE) was also applied to a large set of shooting trajectories at 302 K, in which the trajectories were initiated from an equilibrium distribution within each state. The results of this comparison at 302 K at a lagtime of  $\tau = 6$  ps between the distributions of transition probabilities  $\mathbf{T}(\tau)$  of the different estimation methods are shown in Figure 4.4. Here, the 95%-confidence intervals of the distributions are given in the lower half of each plot. All colors are consistent with Figure 4.3 and Table 4.1.

Overall the reweighting method performs very well compared to the single-temperature estimates. Even transition probabilities that are sampled very poorly at 302 K (such as transitions involving high free energy states 5 and 6) are in good agreement with the reference values at 302 K. Table 4.3 shows the standard deviation in the abso-



**Figure 4.3 – Comparison of all inter-state transition probabilities at a lagtime of  $\tau = 6$  ps as a function of temperature.** Shaded regions denote 95% confidence intervals about the estimated mean. Blue lines show the transition probabilities estimated using dynamical reweighting (DR). Red lines show the estimates from transition matrix estimation (TE) computed from only single temperature data. The black cross-hair indicates the reference using the shooting trajectory data (ST) at 302 K only.

	RMSE at 302 K		
	DR	TE	RE
Low free-energy states (1,2,3,4)	0.007	0.019	0.020
High free-energy states (5,6)	0.140	0.223	0.162
All Transitions	0.079	0.126	0.092

**Table 4.3 – Comparison of error in estimation** Root-mean squared error (RMSE), computed from the absolute difference of transition probabilities for  $\tau = 6$  ps to the reference simulation dataset (ST) at 302 K for the three methods of Markov model estimation and high and low free-energy subsets of transitions.

lute difference of the estimation methods compared to the reference simulation (ST) using a lagtime of  $\tau = 6$  ps. The dynamical reweighting (DR) estimates have a smaller deviation (see also 95%-confidence intervals in Figure 4.4) than both Bayesian methods for both high and low free-energy states.

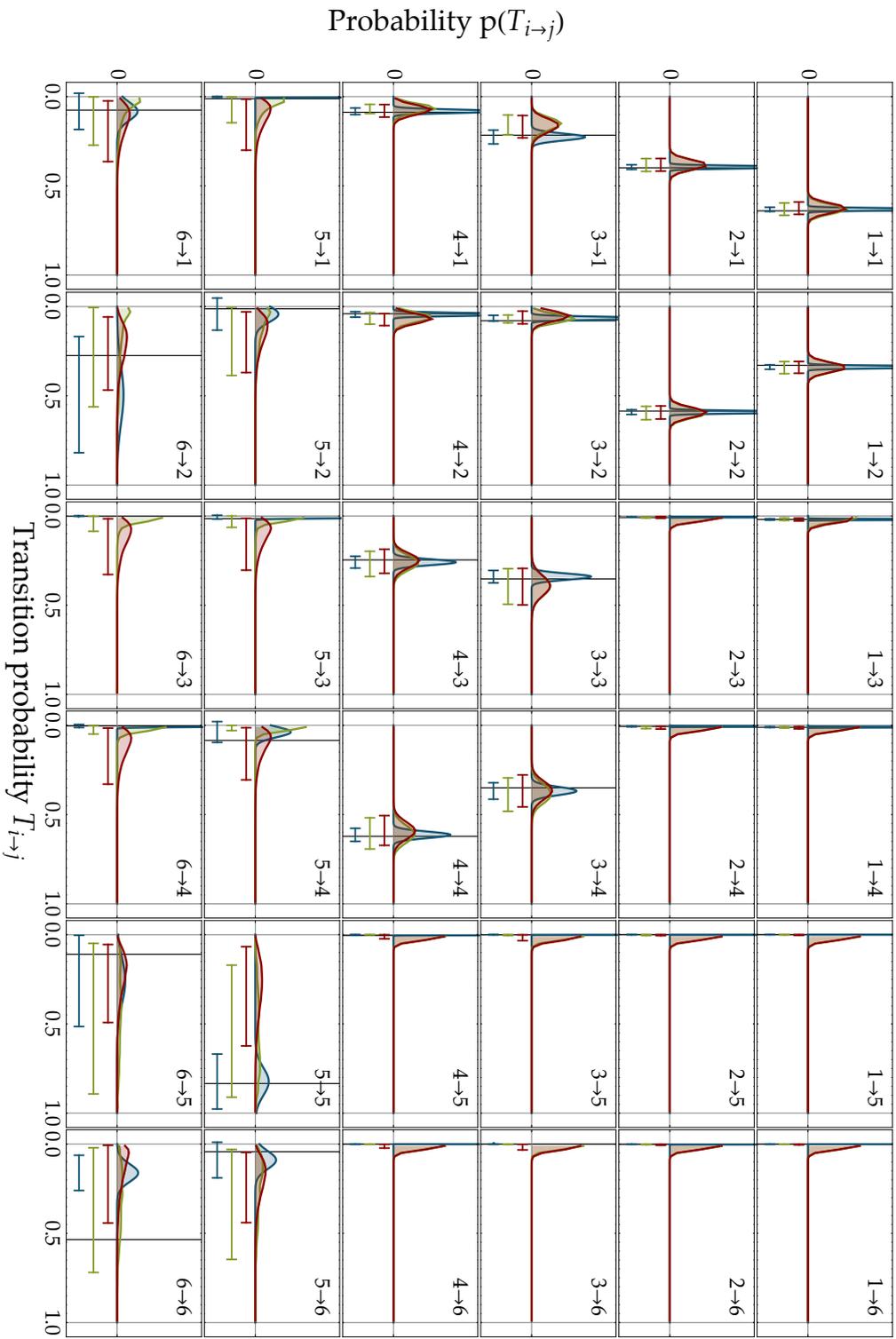
For transitions that are not sampled at certain temperature ranges, the maximum-likelihood estimates obtained with the presented reweighting method are close to zero (see Figure 4.3). Generally, for transition probabilities close to zero or unity, the asymptotic normal distribution assumed in the statistical error estimate of dynamical reweighting is a poor approximation to this highly asymmetric distribution and therefore tends to overestimate the true uncertainties in these cases (see Figure 4.4). Comparing the two Bayesian methods, we find they behave almost identically for transitions among states where many transition counts are observed, but differ for transitions with few transitions (those involving states 5 and 6). Recall that these methods utilize the same likelihood functions but different parameterizations and priors; the influence of this difference is expected to be most prominent when statistics are poor, which is exactly as observed here.

#### 4.4.4 Comparison of temperature dependence of eigenvalues

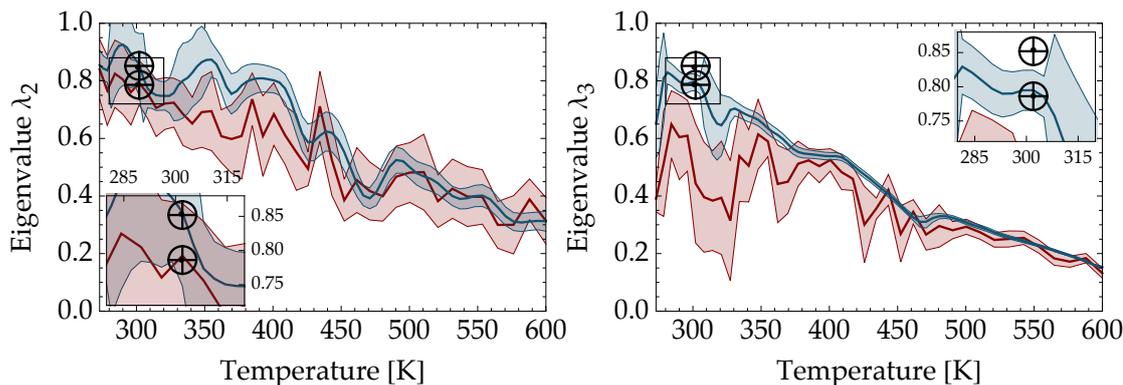
Dynamical reweighting can also be applied to estimate properties derived from the transition probabilities. For example, the eigenvalues  $\lambda_i$  of a transition matrix are related to the time scales of processes  $t_i$  by

$$t_i = -\tau / \ln(\lambda_i), \quad (4.4.1)$$

where we assume that the eigenvalues  $\lambda_i$  are sorted in order of descending absolute value ( $\lambda_1 = 1 > |\lambda_2| > \dots > |\lambda_M| > 0$ ) [79, 53]. Note that Eq. (4.4.1) implies that eigenvalues close to unity are related to slow processes – those we are generally most interested in.



**Figure 4.4 – Detailed comparison of the distribution of transition probabilities  $T(\tau)$  for  $\tau = 6$  ps at 302 K for different estimation methods.**  
 Red: Single temperature estimation of transition matrix (TE); Green: Single temperature estimation of rate matrix (RE); Blue: dynamical reweighting estimation (DR), Black: reference using shooting trajectories (ST). The lower half contains 95%-confidence intervals for each distribution.

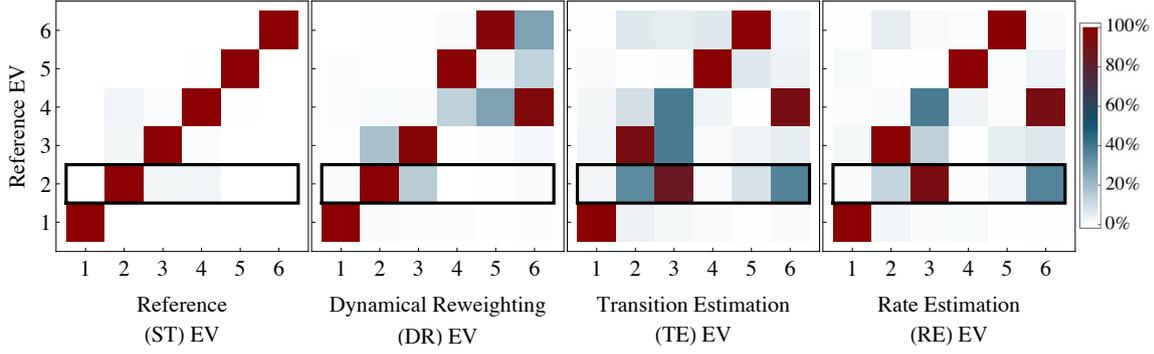


**Figure 4.5 – Temperature dependence of estimated eigenvalues.** Red: Single temperature estimation of transition matrix (TE), Blue: dynamical reweighting estimation (DR), Black: reference 2nd and 3rd eigenvalue at 302 K using shooting trajectories (ST). Upper: Comparison of the second largest eigenvalue vs temperature, Lower: Comparison for third eigenvalue. The third reference eigenvalue is well predicted by both estimation methods at all temperatures although it matches only the second eigenvalue in the transition matrix estimation (TE). The second reference eigenvalue at low temperatures (below 350 K) is only detected by dynamical reweighting (DR).

We investigated the dependence of the eigenvalues on the temperature in the present system. Figure 4.5 compares estimates for the second and third eigenvalues ( $\lambda_2, \lambda_3$ ) of the transition matrix at each temperature with the different methods. The variance in the (TE) case was estimated from the set of eigenvalues of each sampled transition matrix. To estimate the statistical error in the estimates produced by (DR), we used a first-order Taylor expansion to propagate the statistical uncertainties in the transition matrix to uncertainties in the eigenvalues [61].

At low temperatures (below 350 K), the second eigenvalue is estimated correctly by dynamical reweighting (DR), but not by single-temperature estimations. This is due to the fact that the transition process corresponding to this slowest timescale is not sampled at these low temperatures. Thus, estimates using only data collected at that temperature are erroneous. The agreement of dynamical reweighting time scales with the reference simulation is very good, although the error bars of the reweighted estimate are still very large compared to the good agreement of the estimated values with the reference values from the shooting trajectories. We speculate that the inappropriate approximation of the asymmetric posterior distributions with normal distributions used for the linear error propagation may lead here as well to an overestimation of the errors in the transition probabilities.

The third largest eigenvalue is predicted by both methods equally well (Figure 4.5), although it occurs as the second-largest eigenvalue in the single-temperature estimates, which missed sampling the slowest process (described by  $\lambda_2$ ) completely. A direct comparison of the predicted eigenvectors (Figure 4.6) reveals that the slowest process (given by the second eigenvector of the reference transition matrix



**Figure 4.6 – Similarity matrices  $\mathbf{S}$  (scalar product) of eigenvectors from symmetrized transition matrices estimated with different methods at 302 K.** The eigenvectors indicate the states involved in the process, and thus high similarity (red) indicates a good approximation to the reference process (ST). Eigenvectors are sorted as descending eigenvalues. The 2nd eigenvector is found most reliably by dynamical reweighting (DR).

(ST)) is not detected by any of the single temperature methods. However, dynamical reweighting successfully finds all the processes, although the matching eigenvalues, and thus time scales, are permuted for faster processes.

The comparison of Markov models is a nontrivial task [127], for which we use a symmetrized form of the transition matrix  $\mathbf{T}^{\text{sym}}$  and expand it into a sum of rank one matrices spanned by an outer product of the eigenvectors of  $\mathbf{T}^{\text{sym}}$  by

$$\begin{aligned}\mathbf{T}^{\text{sym}} &= \text{diag}((\boldsymbol{\pi})^{1/2}) \mathbf{T} \text{diag}((\boldsymbol{\pi})^{-1/2}) \\ &= \mathbf{R} \text{diag}(\lambda_1, \dots, \lambda_M) \mathbf{R}^T \\ &= \sum_{i \in \mathcal{M}} \lambda_i \mathbf{r}_i \mathbf{r}_i^T.\end{aligned}$$

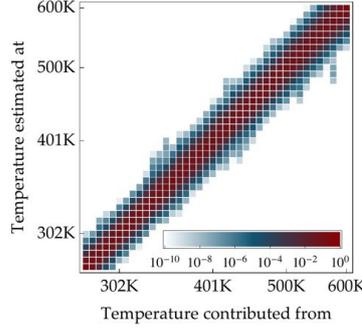
Here,  $\text{diag}((\boldsymbol{\pi})^{1/2})$  means the diagonal matrix with the square root of the equilibrium distribution  $\boldsymbol{\pi}$  on the diagonal,  $\mathbf{R}$  the matrix of normalized eigenvectors of  $\mathbf{T}^{\text{sym}}$  and  $\lambda_i$  the corresponding eigenvalues. Neglecting the time scales (i.e. the eigenvalues  $\lambda_i$ ), two Markov models are similar if their inherent processes, described by the right eigenvectors  $\mathbf{r}_i$ , are similar. Thus, we define the similarity matrix  $\mathbf{S}$  for two transition matrices  $\mathbf{T}$  and  $\mathbf{T}'$  by the mutual scalar product of eigenvectors  $\mathbf{r}_i$  and  $\mathbf{r}'_j$  of their symmetrized form by

$$\mathbf{S} [\mathbf{T} \mid \mathbf{T}'] \equiv \mathbf{R}(\mathbf{R}')^T$$

the results of which are presented in Figure 4.6.

#### 4.4.5 Contributions from different temperatures to the estimates of expectation values

The relative contribution from each temperature to the estimation of any expectation value at a given temperature is presented in Figure 4.7. We plot the relative total contribution from the subset of



**Figure 4.7** – Relative contribution  $\bar{w}_{kl}$  in Eq. (4.4.2) to the estimates at inverse temperature  $\beta_k$  from simulations at inverse temperature  $\beta_l$  summed over all trajectories at the same temperature. On average, seven temperatures contribute more than 1% each to the estimation.

trajectories  $Q_k \subset N = \{1, \dots, N\}$  sampled from the distribution at  $\beta_k$  to estimates at  $\beta_l$  given by

$$\bar{w}(\beta_l | \beta_k) = \frac{\sum_{n \in Q_k} w^{[n]}(\beta_l)}{\sum_{n \in N} w^{[n]}(\beta_l)} \quad (4.4.2)$$

with normalized trajectory weights  $w^{[n]}(\beta)$  as defined in Eq. (4.2.2). On average, seven temperatures contribute more than 1% to the expectation.

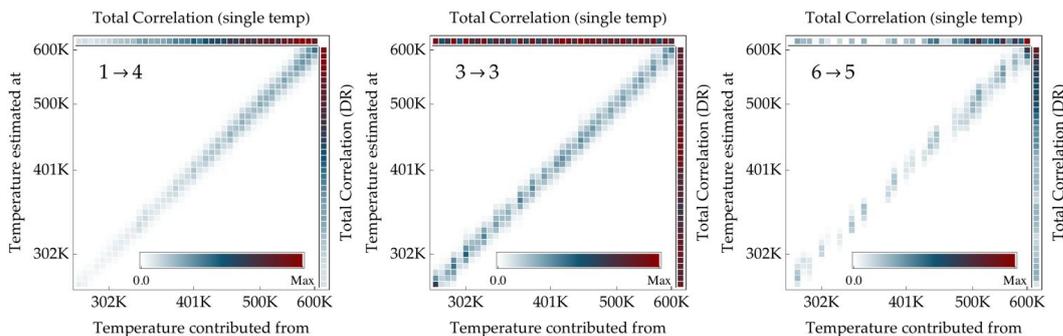
Figure 4.8 illustrates the contribution from the sampled data at  $\beta_k$  to the computed expectation of the time-correlation function (Eq. 4.2.1) for three types of transitions: a highly sampled transition ( $1 \rightarrow 4$ ), a highly sampled remaining in a state  $3 \rightarrow 3$  and a the rarely sampled transition  $6 \rightarrow 5$ . These contributions can be written as

$$\hat{C}_{ij}(\beta_l | \beta_k) = \sum_{n \in Q_k} w^{[n]}(\beta_l) \hat{C}_{ij}^{[n]}, \quad (4.4.3)$$

with  $k$  indicating the temperature contributed from and  $l$  the temperature estimated at. The weighted combination of estimations from multiple temperatures provides a much smoother and continuous estimation (rightmost column) then the estimation from a single temperature (topmost column).

#### 4.5 DISCUSSION

We have presented a method that provides an estimate of Markov state transition probabilities from PT molecular dynamics simulations of biomolecules as a continuous function of temperature. This allows Markov models to be constructed for intermediate temperatures not included in the simulation. In addition, transition matrix estimates at temperatures included in the simulation are much more precise than those obtained with the methods examined here that make use



**Figure 4.8** – Contributions to the estimation of the correlations  $\hat{C}_{ij}$  for the three transition  $1 \rightarrow 4$ ,  $3 \rightarrow 3$  and  $6 \rightarrow 5$  as defined in Eq. (4.4.3). The sum of one row (rightmost column), equal to the total counts estimated by the method at the desired temperature, provides a smoother estimate than the single temperature estimation (topmost column).

of data from a single temperature alone. At low temperatures, even when some transitions are not observed at all at the temperature of interest, corresponding transition probabilities can still be estimated by incorporating dynamical information from higher temperatures without resorting to approximate rate laws such as Arrhenius.

Additionally, the estimates of transition probabilities can be differentiated with respect to the inverse temperature  $\beta$ , because the trajectory weights  $w^{[n]}(\beta)$  in Eq. (4.2.2) are differentiable functions of temperature. This allows, in principle, kinetic and thermodynamic properties that depend on temperature derivatives (e.g., heat capacities) to be computed, provided care is taken in dealing with numerical issues since the trajectory weights  $w^{[n]}(\beta)$  can easily span hundreds of orders of magnitude.

In our illustrative calculations, we chose to employ a modified PT protocol to produce a series of Hamiltonian trajectories with initial configurations drawn from the NVT ensemble. However, the approach itself is not limited to Hamiltonian trajectories, but can be extended to other dynamics (e.g. Brownian and Langevin dynamics) provided a temperature-independent dynamical analogue of the Hamiltonian can be computed, as described in Ref. [4].

The way in which the transition probabilities are estimated in terms of equilibrium correlation functions requires that the trajectory segments sampled during the PT simulation are drawn from equilibrium, which is ensured by the modified PT protocol provided the simulation is sufficiently long (see Appendix B.2). For systems with long correlation times, there may be very few *uncorrelated* trajectories sampled from global equilibrium, but mixing within the Markov states may be sufficiently fast for many uncorrelated phase space configurations to be generated *within* each state. In this case, it is conceivably advantageous to apply dynamical reweighting to the trajectories originating from each set  $\omega_i \subset \Omega$  separately to estimate each row of the transi-

tion matrix  $\mathbf{T}(\tau)$  separately, though the resulting matrix is no longer guaranteed to satisfy detailed balance.

The degree to which the use of PT can enhance the thermodynamic sampling efficiency is limited. Although activated processes will be sampled more often at higher temperatures, it becomes less that entropic bottlenecks will be penetrated. Lower temperatures, on the other hand, lead to an increased sampling of entropic barriers, while at the same time decreasing transitions across enthalpic barriers. This effectively limits the possible improvement in sampling and the range of temperatures that contribute substantial weight to transitions at a given temperature of interest. Despite this, dynamical reweighting allows information about activated processes to be transferred from higher to lower temperatures and, for entropic barriers, from lower to higher temperatures.

Both single-temperature methods give similar results for transitions with good statistics, differing mostly for transitions that were only rarely or not at all sampled, due presumably to the dependence on the choice of a Bayesian prior. The rate matrix estimation (RE) assumes, in addition to the detailed balance constraint present in both methods, positivity of all transition matrix eigenvalues and non-negative off-diagonal rate matrix entries. The uniform distribution of parameters in logarithmic space leads most likely to favoring of low transition probabilities in states with poor transition statistics. Surprisingly, the reversibility constraint seems to enable the Bayesian estimates to provide a reasonable bound on transition probabilities to and from a state even when the state is not sampled at all.

The general predictions of dynamical reweighting are very good, while the quality of the statistical error estimation is limited near extremely small or large transition probabilities due to the reliance on asymptotic normality in the errors. Some combination of Bayesian and reweighting methods (such as T-WHAM [115]) may provide the best of both types of estimators by yielding more accurate uncertainties at the expense of introducing some bias from the introduction of energy histograms or some other parametric distributions for describing the energy density of states (now transitions). Finally, the enhanced estimates of mean values and their respective statistical uncertainties may be used to guide subsequent (potentially adaptive) sampling strategies, as described in Ref. [60].



# 5

---

## ESTIMATION FROM EXPERIMENTAL TIME SERIES

---

This chapter is partially based on the publication

- [8] Prinz, J.-H., Chodera, J. D. & Noé, F. Spectral rate theory for two-state kinetics. *submitted to Physical Review X, ArXiv preprint physics* (2012). url <http://arxiv.org/abs/1207.0225>.

### 5.1 INTRODUCTION

The description of complex molecular motion through simple kinetic rate theories has been a central concern of statistical physics and is the goal of our construction of Markov state models. In contrast to the previous chapters where these simplified models were parametrized from simulations at the atomistic level, experiments concerned with molecular processes often allow only access to low-dimensional projections or are limited to ensemble averages. The recent maturation of measurement techniques that are able to collect extensive traces of single molecule extensions or fluorescence [128, 129] has led to an increased interest in the analysis of these single molecule observations.

One common approach, first-order rate theory, treats the relaxation kinetics among distinct regions of configuration space by single-exponential relaxation. When the available observable or order parameter is a good reaction coordinate that allows the slowly-converting states to be clearly separated, classical rate theories apply and the robust estimation of transition rates is straightforward using a variety of means [130]. However, in the common case in which the reaction coordinate is poor and the separation of the slowly-converting states is not obvious, a satisfactory theoretical description is missing and many estimators break down. In section 2.4.3 the difficulties in constructing Markov models using projections were extensively discussed for the case of various crisp and fuzzy clusterings and lead to error bounds on the approximation quality with respect to the projection. This finally implied enhanced strategies in the choice of projections to minimize this error.

In this chapter we will deal with the problem to (re-)construct key properties of an underlying Markov process from low-dimensional single-molecule projections. This will allow to estimate rates and time

scales with increased accuracy and much improved convergence behavior. The theory is derived with an discussion of systematic and statistical errors and a procedure to apply the new method for the actual estimation of time scales. Later, the method is tested for various sample cases with and without statistical uncertainty. It is finally applied to real experimental trajectories from an optical tweezer experiment of a 56-base single-stranded RNA (ssRNA) hairpin revealing 3 additional relaxation time scales beside the 1 apparent slowest one in a 2-state system. For a more detailed analysis also a physics-based model of the experimental setup. This finally allows to match the experimental findings to actual sets of RNA configurations and interpret the relaxation processes in terms of configurational changes in the hairpin.

## 5.2 CLASSICAL APPROACHES

Our goal is to establish a connection between properties of the original process and the observed time series. At best, the recovery of the dominant time scales and eigenvectors, which define the structural changes occurring at the respective time scale of these processes, would allow for the construction of a simplified model of the relevant underlying kinetics.

There have been a wide variety of approaches to estimate the dominant time scales: Most rate theories and estimators are based upon dividing the observed state space  $Y$  into reactant, product and intermediate substates and then in some way counting transition events that cross the dividing surface. Transition state theory measures the instantaneous flux across this surface, which is known to overestimate the rate due to the counting of unproductive re-crossings over the dividing surface on short time scales [131]. Reactive flux theory [132] has proposed to cope with this by counting a transition event only if it has succeeded to stay on the product side after a sufficiently long lag time  $\tau$ . Reactive flux theory involves derivatives of autocorrelation functions that are numerically unreliable to evaluate [133]. In practice, one therefore typically estimates the relaxation rates *via* single- or multi-exponential fits to a suitable correlation function, such as the number correlation function of reactants or the autocorrelation function of the experimentally measured signal [134, 135, 130].

In the previous chapters it was shown, that especially Markov State Models (MSMs) provide powerful means in producing a simplified statistical model of complex molecular dynamics. These allow the computation of dominant time scales and can thus be regarded as an attempt towards a multi-state rate theory. When the state division allows the metastable states of the system to be distinguished [65, 52, 51, 1], the transition matrix can be used to derive a phenomenological transition rate matrix for sufficiently large lag times

$\tau$  [53]. It has been shown in [49, 1] (see also chapter 2) that by increasing the number of substates used to partition state space, and hence using multiple dividing surfaces instead of a single one, these rate estimates become more precise. In the limit of infinitely many substates, the eigenfunctions of the dynamical propagator  $T$  in the full phase space are exactly recovered, and the rate estimates become exact even for  $\tau \rightarrow 0^+$  [75]. In practice, however, a finite choice of  $\tau$  is unavoidable due to the finite sampling rate and necessary in order to have a small systematic estimation error, which is caused by the projection onto the experimental observables. An alternative way of estimating transition rates is by using a core based state definition that is incomplete and treats the transition region implicitly *via* the committor functions from chapter 3 [44, 108].

The quality of the rate estimates in all of the above approaches (MSM based or not) relies on the ability to separate the slowly-converging states in terms of some dividing surface or fuzzy state definition. The better the state definition the smaller the implied error in the estimated rates. In the case where the available observables or order parameter(s) do not permit such a separation, i.e. when kinetically distinct states overlap in the histogram of the observed order parameter, most estimators break down or fail to provide estimates within a specified confidence interval.

In data collected from simulations the full phase-space information is present and one has wide freedom to optimize the simplifying projections. However, in single-molecule experiments where the available observation or order parameter is limited by the experimental setup, this is usually not the case and there might be no choice of projecting the observation to be a good reaction coordinate. While Hidden Markov Models (HMMs) are able to estimate transition rates even in such situations [136, 137, 138], they depend on a probability model of the measurement process and rather represent a computational optimization approach than a physically motivated theory. Also, HMMs are relatively complex to implement and the often used estimation-maximization (EM) algorithms are based on local optimization which requires a check of convergence.

### 5.3 OBSERVATIONS AND PROJECTIONS

In the following we will assume that the underlying dynamics can be exactly described by a (stochastic) reversible Markov process  $X_t$ <sup>1</sup> on some state space  $\Omega$  as we have defined in chapter 2 and thus all information of the dynamics is contained in the transfer operator  $T$  or, in an approximated discretized context, in the transition matrix  $\mathbf{T}$ .

<sup>1</sup> Here,  $X_t$  represents the Markov process itself while  $x_t$  refers to a concrete realization of the original process in the full state space  $\Omega$ .

### 5.3.1 Observed Dynamics

In most experiments the underlying Markovian dynamics is only partially observable. This is also the case in simulations if not the full phase space (including solvent and velocities) is recorded. Hence, a mathematical object is necessary that can describe the relation between a certain state in the full-dimensional phase space  $\Omega$  and its representation in the *projected* or *observable* space  $Y$ . We chose this to be real-valued,  $Y \subset \mathbb{R}$ , to be in accordance with a physically measurable quantity and call a surjective function

$$\zeta : x \in \Omega \mapsto \zeta(x) \in Y \subset \mathbb{R}$$

that maps between these two spaces a *projection*  $\zeta$ . To deal with higher-dimensional observed spaces,  $Y \subset \mathbb{R}^d$ , the function can be split into a set  $\xi = \{\xi_1, \dots, \xi_d\}$  of single projections while complex-valued functions can be treated by splitting them into real and imaginary parts.

While in a full atomic molecular simulations arbitrary projections can be chosen, this is not possible in an experiment: The concrete setup of the experiment determines the measurable output, which can then be analyzed further. We will call projections that are determined by an experimental setup or can be measured by other means an *observable*  $y(x)$ . These observables are *a priori* given and if applied to a time series  $x_t \in \Omega$ , the experimentally observed time series  $y_t \equiv y(x_t)$  is recovered. If the available projections that will later be used in the analysis are to be limited by this fixed observable  $y$ , we can express this by  $\zeta(x) = \zeta(y(x))$  and regarded this as a two-fold projection: The first (inner) projection is fixed by the experiment and the second can be chosen freely by an algorithm (see Figure 5.1).

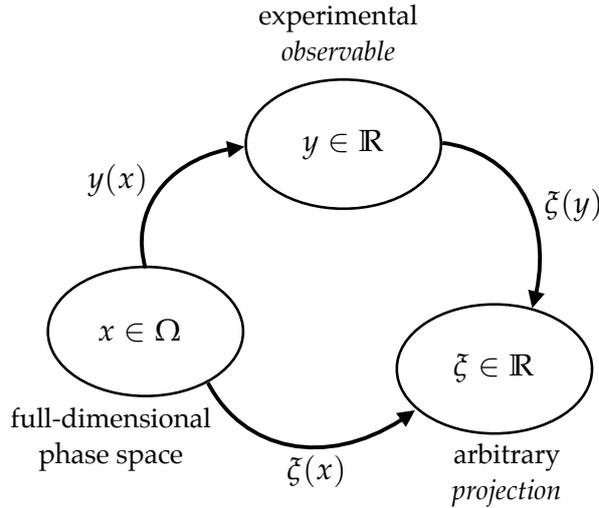
To clarify the terminology: All functions  $\zeta$  from  $\Omega$  to  $\mathbb{R}$  are referred to as projections while observations  $y$  are a special type of projections, that are directly accessible by an experiment or simulation.

### 5.3.2 Correlations

To proceed, we first introduce the *correlation density*  $c(x, y)$  that contains the absolute probability density for a transition from  $x \in \Omega$  to  $y \in \Omega$  to occur according to the given dynamics. The density is defined by

$$\begin{aligned} c_\tau(x, A) &\equiv \pi(x) p_\tau(x, A) \\ &= \mathbb{P}[x_t = x] \mathbb{P}[x_{t+\tau} \in A | x_t = x] \\ &= \mathbb{P}[x_{t+\tau} \in A, x_t = x], \end{aligned}$$

the product of the conditional jump probability density  $p_\tau(x, y)$  weighted with the stationary probability density  $\pi(x)$  to find the system



**Figure 5.1 – Functional relation between projections and observations** Observations  $y$  are projections from the original full phase space  $\Omega$  that can be observed by an experiment or simulation. For further analysis these observed quantities are often projected onto a specific and problem specific general projection  $\zeta$  that itself can also be seen as a two-fold projection.

initially in a state  $x \in \Omega$ . In the case of microscopic reversibility,  $c_\tau$  is symmetric

$$c_\tau(x, y) = c_\tau(y, x)$$

and it therefore exists a symmetric representation in terms of eigenfunctions  $\psi_i$  and eigenvalues  $\lambda_i, i \in I$  expressed as

$$c_\tau(x, y) = \sum_{i \in I} \lambda_i^\tau \pi(x) \psi_i(x) \pi(y) \psi_i(y) \tag{5.3.1}$$

using that the eigenfunctions have been chosen  $\psi_i(x)$  are orthonormal w.r.t. the  $\pi$ -weighted scalar product

$$\int dx \pi(x) \psi_i(x) \psi_j(x) = \langle \psi_i(x), \psi_j(x) \rangle = \delta_{ij}.$$

As before,  $I$  is the countable index set for the eigendecomposition to allow for both, finite  $I = \{1, \dots, M\}$  and infinite  $I = \{1, \dots, \infty\}$  numbers of eigenvector/eigenvalue pairs. In both case the elements are considered to be ordered by the decreasing modulus of the eigenvalue.

Assuming that we have access to arbitrary projections, one can recover the eigenvalues of the transition operator  $T$  from the original jump process  $X_t$  if the observation is projected onto an eigenfunction  $\psi_i(x)$ . For an arbitrary projection  $\zeta$  the autocorrelation  $\text{acf}[\cdot]$  of  $\zeta(x_t)$  for a trajectory  $x_t$  of length  $L$  is an expectation  $\langle \cdot \rangle_t$  over the time  $t$  and can be written as

$$\begin{aligned} \text{acf}[\tilde{\zeta}(x_t)](\tau) &= \langle \tilde{\zeta}(x_t) \tilde{\zeta}(x_{t+\tau}) \rangle_{t \in [0, L-\tau]} \\ &= (L-\tau)^{-1} \int_0^{L-\tau} \mathbf{d}t \tilde{\zeta}(x_t) \tilde{\zeta}(x_{t+\tau}) \end{aligned}$$

and we extend the product  $\tilde{\zeta}(x_t) \tilde{\zeta}(x_{t+\tau})$  by two integrals of unity that yield

$$\begin{aligned} \tilde{\zeta}(x_t) \tilde{\zeta}(x_{t+\tau}) &= \iint_{\Omega^2} \mathbf{d}a \mathbf{d}b \delta(a - x_t) \delta(b - x_{t+\tau}) \tilde{\zeta}(x_t) \tilde{\zeta}(x_{t+\tau}) \\ &= \iint_{\Omega^2} \mathbf{d}a \mathbf{d}b \delta(a - x_t) \delta(b - x_{t+\tau}) \tilde{\zeta}(a) \tilde{\zeta}(b). \end{aligned}$$

If we now define the estimated correlation density for transitions between  $a, b \in \Omega$ , similar to the estimated correlation for discrete time series, as

$$\begin{aligned} \hat{c}_\tau(a, b | L) &= (L-\tau)^{-1} \int_0^{L-\tau} \delta(a - x_t) \delta(b - x_{t+\tau}) \\ &= \mathbb{P}(x_{t+\tau} = b | x_t = a) \end{aligned}$$

we can rewrite the autocorrelation in terms of transitions between all possible projected values in the full state space  $\Omega$  by

$$\text{acf}[\tilde{\zeta}(x_t)](\tau) = \iint_{\Omega^2} \mathbf{d}a \mathbf{d}b \hat{c}_\tau(a, b | L) \tilde{\zeta}(a) \tilde{\zeta}(b).$$

If the time series is infinitely long ( $l \rightarrow \infty$ , i. e. , we can estimate without statistical uncertainty) the estimated correlation density  $\hat{c}_\tau$  converges to the true correlation density

$$\lim_{L \rightarrow \infty} \hat{c}_\tau(a, b | L) = c_\tau(a, b)$$

and we find that in this limit the autocorrelation is equally represented by an estimation over time

$$\begin{aligned} \text{acf}[\tilde{\zeta}(x_t)](\tau) &= \langle \tilde{\zeta}(x_t) \tilde{\zeta}(x_{t+\tau}) \rangle_{t \in \mathbb{R}^+} \\ &= \langle \tilde{\zeta}(a) c_\tau(a, b) \tilde{\zeta}(b) \rangle_{a, b \in \Omega} \end{aligned}$$

as well as an estimation over all possible transitions  $\langle \rangle_{a, b \in \Omega}$ . If we now have an intermediate experimental observation  $y$  and our projection only depends on the observed trajectory  $y_t$  we can use the same argument and derive

$$\begin{aligned} \text{acf}[\tilde{\zeta}(y_t)](\tau) &= \langle \tilde{\zeta}(y_t) \tilde{\zeta}(y_{t+\tau}) \rangle_{t \in \mathbb{R}^+} \\ &= \langle \tilde{\zeta}(a) c_\tau^Y(a, b) \tilde{\zeta}(b) \rangle_{a, b \in \mathbb{R}} \end{aligned}$$

where we have now to use the *projected* correlation density

$$c_\tau^Y(a, b) = \mathbb{P}(y_{t+\tau} = b | y_t = a).$$

Note, that a Markov model constructed from this observed correlation density will not fulfill the Chapman-Kolmogorov semi-group property as was explained in chapter 2.4.

Using the expansion for the correlation density  $c_\tau(x, y)$  in Eq. (5.3.1) we derive

$$\begin{aligned}
 \text{acf}[\tilde{\zeta}(x_t)](\tau) &= \langle \tilde{\zeta}(a) c_\tau(a, b) \tilde{\zeta}(b) \rangle_{a, b \in \Omega} \\
 &= \langle \tilde{\zeta}(a) \sum_{i \in I} \lambda_i^\tau \pi(a) \psi_i(a) \pi(b) \psi_i(b) \tilde{\zeta}(b) \rangle_{a, b \in \Omega} \\
 &= \sum_{i \in I} \lambda_i^\tau \langle \tilde{\zeta}(a) \pi(a) \psi_i(a) \pi(b) \psi_i(b) \tilde{\zeta}(b) \rangle_{a, b \in \Omega} \\
 &= \sum_{i \in I} \lambda_i^\tau \langle \tilde{\zeta}(a), \psi_i(a) \rangle_{a \in \Omega} \langle \psi_i(b), \tilde{\zeta}(b) \rangle_{b \in \Omega} \\
 &= \sum_{i \in I} \lambda_i^\tau (\langle \tilde{\zeta}(a), \psi_i(a) \rangle_{a \in \Omega})^2
 \end{aligned}$$

so that the autocorrelation  $\text{acf}[\tilde{\zeta}(x_t)](\tau)$  can always be expressed as a sum of exponential decays which is the origin of the exponential fitting in classic rate theory. If now the projection  $\tilde{\zeta}(x)$  can be chosen to be an eigenfunction  $\psi_k$  we compute

$$\begin{aligned}
 \text{acf}[\psi_k(x_t)](\tau) &= \sum_{i \in I} \lambda_i^\tau \langle \psi_k(a), \psi_i(a) \rangle^2 \\
 &= \sum_{i \in I} \lambda_i^\tau (\delta_{ik})^2 \\
 &= \lambda_k^\tau
 \end{aligned} \tag{5.3.2}$$

and recover a single-exponential decay with the original time scale  $\lambda_k$  where the orthonormality of the eigenfunctions was used. Thus we can reconstruct  $\lambda_k$  by a single-exponential fit to this particular autocorrelation function.

The necessary exact eigenfunctions are usually only approximately (if at all) accessible, but the last formulation suggests an expansion of the projection  $\tilde{\zeta}$  into eigenfunctions of the full dynamics

$$\begin{aligned}
 \tilde{\zeta}(x) &= \sum_{i \in I} \langle \psi_i, \tilde{\zeta} \rangle \psi_i(x) \\
 &= \sum_{i \in I} q_i \psi_i(x)
 \end{aligned}$$

with  $q_i \equiv \langle \psi_i, \tilde{\zeta} \rangle$ ,  $i \in I$  measuring the *overlap* between the projection  $\tilde{\zeta}$  and exact eigenfunction  $\psi_i$ . If the projection is normalized we can use Parseval's identity and

$$\begin{aligned}
 1 = \langle \tilde{\zeta}, \tilde{\zeta} \rangle &= \sum_{i, j \in I} \langle q_i \psi_i(x), q_j \psi_j(x) \rangle \\
 &= \sum_{i, j \in I} q_i q_j \langle \psi_i(x), \psi_j(x) \rangle \\
 &= \sum_{i, j \in I} q_i q_j \delta_{ij} \\
 &= \sum_{i \in I} q_i^2
 \end{aligned}$$

follows, so that the vector  $\mathbf{q} \equiv \{q_i \mid i \in I\}$  is of unit length in the euclidean norm  $\|\mathbf{q}\|_2 = 1$ . This  $\mathbf{q}$  can conveniently be used to measure the quality of a projection  $\zeta$  to approximate a particular eigenfunction  $\psi_k$ : If and only if one single eigenvector  $\psi_k$  is present in the observation then the corresponding  $q_k$  equals one. Conversely, if the maximum of  $\mathbf{q}$  is smaller than one (in particular  $q_k$ ), more than one eigenvector must be present.

In the following we will assume that any projection  $\zeta$  (independent of being based on an actual observation or not) used is normalized  $\langle \zeta, \zeta \rangle = 1$  and chosen to be orthogonal to the stationary process  $\langle \zeta, \psi_1 \rangle = 0$ . Orthogonality can always be achieved by adding a constant function

$$\mathbf{k} = -\mathbf{1} \cdot \langle \zeta, \psi_1 \rangle$$

to the observed values

$$\begin{aligned} \langle \zeta + \mathbf{k}, \psi_1 \rangle &= \langle \zeta, \psi_1 \rangle + \langle \mathbf{k}, \psi_1 \rangle \\ &= \langle \zeta, \psi_1 \rangle - \langle \mathbf{1} \cdot \langle \zeta, \psi_1 \rangle, \psi_1 \rangle \\ &= \langle \zeta, \psi_1 \rangle - \langle \zeta, \psi_1 \rangle \langle \mathbf{1}, \psi_1 \rangle \\ &= \langle \zeta, \psi_1 \rangle - \langle \zeta, \psi_1 \rangle = 0 \end{aligned}$$

since the eigenfunction of the stationary process is constant. In this case the correlation takes the form

$$\text{acf}[\zeta(x_t)](\tau) = \sum_{i \geq 2} \lambda_i^\tau q_i^2$$

which implies that the exponential decays in the correlation function are now determined by two factors:

1. The overlap of the projection with the eigenvectors given by  $q_i$ : The closer the maximum of  $\mathbf{q}$  is to one the more single-exponential is the auto-correlation. The idea of optimization and analysis of these  $q$ -factors is extensively used in dynamical fingerprints [139, 6]
2. The time scales  $\lambda_i$  and the chosen lag time  $\tau$ . With increasing lag time  $\tau$ , more and more exponentials decay to zero as was explained in chapter 2.2.

Hence, to estimate time scales by a single exponential, one can either optimize the used observation  $\zeta$  to match the corresponding eigenfunction  $\psi_i$  or, if only the slowest process is of interest, increase the lagtime until only this slowest process, given by  $\psi_2(x)$ , remains.

5.3.3 Correlation Matrices

The natural generalization of the autocorrelation is to compute cross-correlations between different observations  $\xi_i, \xi_j, i, j \in \{1, \dots, n\}$  which are then defined as

$$\begin{aligned} \text{cor}[\xi_i(x_t) | \xi_j(x_t)](\tau) &= \langle \xi_i(x_t) \xi_j(x_{t+\tau}) \rangle_t \\ &= \langle \xi_i(a) c_\tau(a, b) \xi_j(b) \rangle_{a,b \in \Omega} \end{aligned}$$

and these cross-correlation functions can be conveniently written as

$$\text{cor}[\xi_i(x_t), | \xi_j(x_t)](\tau) = \sum_{k \neq 1 \in I} \lambda_k^\tau Q_{ki} Q_{kj}$$

with

$$Q_{ij} = \langle \psi_i, \xi_j \rangle$$

being the matrix of projected eigenvectors  $\mathbf{Q}$ . We finally define a  $\tau$ -dependent projected correlation matrix  $\mathbf{C}^X(\tau)$  that is induced by the set  $\xi = \{\xi_1, \dots, \xi_n\}$  of projections by

$$\begin{aligned} \mathbf{C}^X(\tau) &\equiv \text{cor}[\xi(x_t) | \xi(x_t)](\tau) \\ &= \langle c_\tau(a, b) \xi(a) \xi^\top(b) \rangle_{a,b \in \Omega} \end{aligned} \tag{5.3.3}$$

where the expectation is computed over the outer product  $\xi(a) \xi^\top(b)$  so that

$$C^X(\tau)_{ij} = \text{cor}[\xi_i(x_t) | \xi_j(x_t)](\tau)$$

holds.

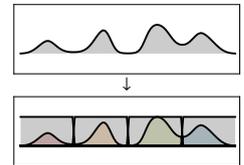
5.3.4 Special Types of Projections

There exists a variety of projections with a special meaning or area of application, which we want to present in the following. Note, that all scenarios are in their mathematical treatment fully equivalent while only the meaning or interpretation of the projection changes.

*Crisp Clustering*

High-dimensional (continuous) time series usually need to be reduced to a finite state space to be analyzed computationally (see chapter 2). To achieve this, the state space  $\Omega$  is projected on a finite number of macro states. For such a crisp decomposition defined by the subsets of the full state space  $\omega_i \subset \Omega, i \in M$  with mutual exclusive states  $\omega_i \cap \omega_j = \emptyset, i \neq j$  the associated projection is given by

$$\xi_i(x) = \chi_i(x) = \begin{cases} 1 & \text{if } x \in \omega_i \\ 0 & \text{else} \end{cases} \tag{5.3.4}$$



**Example**  
Crisp Clustering

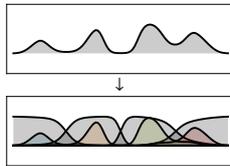
which is a simple indicator function for the subsets  $\omega_i$  that measure if state  $x \in \Omega$  belongs to subset  $\omega_i$ . The resulting observed correlation matrix  $\mathbf{C}^X(\tau)$  now contains the absolute probabilities of transitions between subsets in the clustered system which we have seen to be used construct Markov State Models, approximating the observed dynamics in section 2.5.

A more general approach is to loosen the constraint of mutual exclusivity for the clusters and merely demand, that the complete space is contained in the decomposition much as in a partition of unity. The observations  $\xi$  now measure the (non-negative) percentage that a micro state  $x \in \Omega$  belongs to a given macro state  $i \in M$ , which has to add up to unity

$$\xi_i(x) = \chi_i(x) \in [0, 1], \sum_i \chi_i(x) = 1, \forall x \in \Omega. \tag{5.3.5}$$

The resulting observed correlation matrix again contains absolute probabilities to make a transition within this projected system. In the case of a crisp or fuzzy clustering we may also call the induced projections  $\chi_i$  a *clustering*.

*Linear Combinations of Eigenfunctions (PCCA)*



*Example  
PCCA Clustering*

As discussed extensively in the works by Röblitz, Weber and Deuffhard [43] there exists a very special case, where the full dynamics is projected into a subspace spanned by the  $m$  dominant eigenvectors of the transfer operator or transition matrix. Then the projection error discussed in chapter 2 vanishes and the projected transition matrix is again exact in the sense that it evolves the observed projection correctly. The observation in this case takes the form

$$\xi_i(x) = \sum_{j=1}^m A_{ij} \psi_j(x)$$

which represent a linear combination  $\mathbf{A} \in \mathbb{R}^{m \times m}$  of the dominant eigenvectors  $\psi_i$ . Note, that these are usually not accessible, or only

with great numerical effort. We can now compute the observed correlation matrix to be

$$\begin{aligned}
C_{ij}^X(\tau) &= \langle \xi_i(x_t) \xi_j(x_{t+\tau}) \rangle \\
&= \langle \left( \sum_{k=1}^m A_{ik} \psi_k(x_t) \right) \left( \sum_{l=1}^m A_{jl} \psi_l(x_{t+\tau}) \right) \rangle \\
&= \sum_{k,l=1}^m A_{ik} A_{jl} \langle \psi_k(x_t) \psi_l(x_{t+\tau}) \rangle \\
&= \sum_{k,l=1}^m A_{ik} A_{jl} \delta_{kl} \lambda_k(\tau) \\
&= \sum_{k=1}^m A_{ik} A_{jk} \lambda_k(\tau)
\end{aligned}$$

and find

$$\mathbf{C}^X(\tau) = \mathbf{A} \text{diag}(\lambda_1, \dots, \lambda_m) \mathbf{A}^T$$

in matrix notation. Since the (right) eigenvectors  $\psi_i$  do not form a (non-negative) membership, the observations are not strictly positive as well, as would be desired for the interpretation as a probability distribution in a Markov model. This can be fixed by using a suitable linear combination  $\mathbf{A}$  of the eigenvectors that fulfill the requirements of a membership/crisp clustering. This is always possible since the constant vector (i.e. the eigenvector  $\psi_1$ ) is contained in the projection. The actual optimization to find such a suitable membership out of known dominant (right) eigenvectors is called Perron Cluster Cluster Analysis (PCCA) [43] and it can be proven that this is an optimal projection in the sense that the in section 2.4 defined projection error is minimized.

#### 5.4 2-STATE DYNAMICS

If we assume that the full dynamics can approximately be described by the transition between only 2 dominant states, we talk about a 2-state system. According to the results from chapter 2 this can be achieved if there exists a spectral gap between the second and third eigenvalue of the transition matrix. This, in turn, would allow to choose a lag time  $\tau$  so that all, except for the slowest process, have already decayed and the estimation becomes particularly simple. It can thus serve as an introduction to the general multi-state case.

In the following we explicitly assume, that the full dynamics can be approximated to some degree by the stationary process and one further dominant eigenvector  $\psi_2$ . We are now interested in the associated dominant eigenvalue  $\lambda_2$  or the rate  $\kappa_2$  between the two states A and B. Since the necessary eigenfunction  $\psi_2$  is not accessible, the autocorrelation in Eq. (5.3.2) cannot be used to estimate the exact time

scale  $\lambda_2$ . At best, the observation  $\xi$  is close to the second eigenfunction and the expansion into eigenfunctions becomes

$$\begin{aligned}\xi(x) &= q_2\psi_2(x) + \sum_{i>2} \langle \xi, \psi_i \rangle \psi_i \\ &= q_2\psi_2(x) + \epsilon(x),\end{aligned}\quad (5.4.1)$$

where the true eigenfunction  $\psi_2$  enters with coefficient  $q_2$  while the remaining eigenfunctions contribute to the error  $\epsilon(x)$ . The autocorrelation now evaluates to

$$\begin{aligned}\tilde{\lambda}_2(\tau) &= \langle \xi(x_t)\xi(x_{t+\tau}) \rangle \\ &= \alpha\lambda_2^\tau + \sum_{i>2} q_i^2\lambda_i^\tau\end{aligned}\quad (5.4.2)$$

and yields some approximation  $\tilde{\lambda}_2(\tau)$  to the exact eigenvalue, which depends on  $\alpha \equiv q_2^2$ . For better comparison with existing methods, we will use rates

$$\kappa_i = -\tau^{-1} \ln(\lambda_i)$$

instead of eigenvalues in this section and find

$$\exp(-\tau\hat{\kappa}_2) = \tilde{\lambda}_2(\tau) = \alpha \exp(-\tau\kappa_2) + \sum_{i>2} q_i^2 \exp(-\tau\kappa_i) \quad (5.4.3)$$

for the estimated rate  $\hat{\kappa}$ . For  $\tau \gg \kappa_3^{-1}$ , which can easily be achieved for clear 2-state processes where  $\kappa_2 \ll \kappa_3$ , the sum on the right hand side disappears

$$\langle \xi(x_t)\xi(x_{t+\tau}) \rangle \approx \alpha e^{-\kappa_2\tau} \quad (5.4.4)$$

and the autocorrelation is approximately single-exponential with an amplitude given by  $\alpha$ .

#### 5.4.1 Best reaction coordinate

Under the constraint that only a specified experimental order parameter  $y$  is observable the question remains, what the best possible choice for the observable  $\hat{\xi}(y(x))$  is? Following Eq. (5.4.3), the optimal choice  $\hat{\xi}$  is the one which maximizes the parameter  $\alpha$ , as this will minimize the systematic error for the proposed estimation scheme. We are thus seeking the solution of

$$\hat{\xi} = \arg \max_{\xi} \alpha(\xi) \quad (5.4.5)$$

or equivalently

$$\begin{aligned}\hat{\xi} &= \arg \max_{\xi} \tilde{\lambda}_2(\tau; \xi) \\ &= \arg \max_{\xi} \langle \xi(x_t)\xi(x_{t+\tau}) \rangle\end{aligned}$$

subject to the normalizations that  $\langle \zeta, \zeta \rangle = 1$  and  $\langle \zeta, \mathbf{1} \rangle = 0$ . Here,  $\arg \max_{\zeta} \alpha$  denotes the function that maximizes  $\alpha$  over the space of functions  $\zeta(y(x))$  that are usually induced by the experimental observable  $y(x)$ . If the optimal eigenfunction  $\hat{\zeta}$  is supposed to be approximated by a linear combination

$$\hat{\zeta}(y) \approx \sum_{i=1}^n c_i g_i(y)$$

with coefficients  $c_i$  over a set of  $n$  basis functions,

$$\mathcal{G} = \{g_1(y), \dots, g_n(y)\},$$

then the optimal observable  $\hat{\zeta} = \arg \max_{\zeta} \alpha(\zeta)$  can be approximated on  $\mathcal{G}$  using the Ritz method [140]. Since the optimization still depends on the given observable  $y$ , which can be regarded as the experimental reaction coordinate, so will the actual maximal value of  $\alpha$ . If we recall that the  $\alpha$  measures how single-exponential the projection is this can be used to investigate the quality of the observation.

#### 5.4.2 Reaction coordinate quality (RCQ)

Evaluating how well a given putative reaction coordinate captures complex dynamical behavior is of great general interest. Previous studies have proposed ways to measure the Reaction Coordinate Quality (RCQ) that are based on comparing the observed dynamics to specific dynamical models or testing the ability of the observable to model the committor or splitting probability between two chosen end-states A and B [141]. These metrics are either only valid for specific models of dynamics or themselves require a sufficiently good separation of A and B by definition, restricting their applicability to observables with rather good RCQs.

The pre-factor  $\alpha$  in Eq. (5.4.4) is a measure between 0 and 1 quantifying how good a reaction coordinate the observation function  $\zeta(y(x))$  is, and, as will be shown later by virtue of Eqs. (5.5.2) and (5.5.3), how large the error in our rate estimate can be. This RCQ  $\alpha$  can be estimated by an exponential fit to the autocorrelation function (5.4.4) of the observation function  $\zeta$  employed. For the best possible choice  $\zeta = \hat{\zeta}$  (Eq. (5.4.5)), we denote this pre-factor  $\hat{\alpha}$ , where  $\hat{\alpha} = \alpha(\hat{\zeta}) \geq \alpha(\zeta)$ , i.e.  $\hat{\alpha}$  is the best possible (maximal) RCQ of the observable used under a given experimental observation  $y(x)$ . Thus, we propose  $\hat{\alpha}$  as a general measure for the quality of the reaction coordinate for the later discussed classes of estimators that are based on fits to the autocorrelation.

This RCQ is as general as possible, as it makes no assumptions about the class of dynamics in the observed coordinate, and does not depend on any subjective choices such as the choice of two reaction end-states A and B in terms of the observable  $y$ . Through the

derivation above it has also been shown that  $\hat{\alpha}$  measures the fraction of amplitude by which the slowest process is observable, which is exactly the property one would expect from a measure of the RCQ:  $\hat{\alpha}$  is 1 for a perfect reaction coordinate and 0 if the slowest process is exactly orthogonal to the observable.

For completeness one should note, that the newly defined  $\alpha$  is directly related to the projection error  $\delta_2$  of the second eigenvector defined in Eq. (2.4.5) in section 2.4.3 by

$$\alpha = 1 - \delta_2^2 = 1 - \sum_{i>2} q_i^2$$

as it must since the RCQ  $\alpha$  measures the overlap of the observation  $\zeta$  with the second eigenvector  $\psi_2$ .

## 5.5 ESTIMATION

### 5.5.1 Single- $\tau$ rate estimators

Many 2-state rate estimators can be shown to effectively calculate autocorrelation functions of some function  $\zeta$  at a single value of  $\tau$ , and transforming it into a rate estimate by virtue of Eq. (5.4.4). We call these estimators *single- $\tau$  estimators*. Ignoring statistical uncertainties, they yield a rate estimate of the form

$$\begin{aligned} \hat{\kappa}_2 &\equiv -\tau^{-1} \ln \tilde{\lambda}_2(\tau) \\ &= -\tau^{-1} \ln \left( \alpha \lambda_2^\tau + \sum_{i>2} q_i^2 \lambda_i^\tau \right) \\ &= \kappa_2 - \tau^{-1} \ln \left( \alpha + \sum_{i>2} q_i^2 \exp(-\tau(\kappa_i - \kappa_2)) \right). \end{aligned} \quad (5.5.1)$$

From  $\sum_i q_i^2 = 1$  and  $\kappa_i \geq \kappa_2$ , it follows that  $\hat{\kappa}_2 \geq \kappa_2$ , i.e. the estimated rate  $\hat{\kappa}_2$  always overestimates the true rate  $\kappa_2$ . In general the systematic error in the rate  $\Delta\kappa_2 = \hat{\kappa}_2 - \kappa_2$  can be bounded from above by

$$\Delta\kappa_2 \leq \tau^{-1} \ln \alpha^{-1}$$

as is proven in the appendix C.3. When the system is characterized by a time scale gap  $\kappa_3 \gg \kappa_2$ , the error becomes

$$\hat{\kappa}_2 - \kappa_2 \lesssim \tau^{-1} \ln \alpha^{-1} \quad (5.5.2)$$

which is dominated by a relatively slow  $\tau^{-1}$ -decay in the lag time (see Figure 5.2 for a 2-state example). It will be shown below that methods that estimate rates from counting the number of transitions across a dividing surface, such as Markov state models, are single- $\tau$  estimators and are thus subject to the error given by Eq. (5.5.2).

The systematic error of single- $\tau$  estimators results from the fact that Eq. (5.5.1) effectively attempts to fit the tail of a multi-exponential

decay  $\tilde{\lambda}_2(\tau)$  by a single-exponential with the additional constraint  $\tilde{\lambda}_2(0) = 1$ . Unfortunately, the ability to improve these estimators by simply increasing  $\tau$  is limited because the statistical uncertainty of estimating Eq. (5.4.4) quickly grows in  $\tau$  [142].

### 5.5.2 Multi- $\tau$ rate estimators

To avoid the slow convergence with  $\tau$  of the error given by Eq. (5.5.2), it is advisable to estimate the rate by evaluating the autocorrelation function  $\tilde{\lambda}_2(\tau)$  at multiple values of  $\tau$ . This can be done e.g. by performing an exponential fit to the *tail* of the  $\tilde{\lambda}_2(\tau)$  or its time integral, thus avoiding the constraint  $\tilde{\lambda}_2(0) = 1$  [130, 135]. The estimation error  $\Delta\kappa_2$  depends on the specific algorithm used to fit  $\tilde{\lambda}_2(\tau)$ . Using a simple least-squares procedure at the equally spaced time points  $\tau < 2 \cdot \tau < \dots < m \cdot \tau$  (see appendix C.3), we obtain again an overestimation of the true rate,  $\hat{\kappa}_2 \geq \kappa_2$ , and the systematic estimation error of  $\hat{\kappa}_2$  can be now bounded by

$$0 \leq \Delta\kappa_2 \leq \frac{(m+1)(\bar{\tau} - \tau_1)}{m \text{Var}(\tau)} \ln \left( 1 + \frac{(1-\alpha)}{\alpha} e^{-\tau_1(\kappa_3 - \kappa_2)} \right) \quad (5.5.3)$$

$$< \text{const} \cdot \tau^{-1} \frac{(1-\alpha)}{\alpha} e^{-\tau_1(\kappa_3 - \kappa_2)} \quad (5.5.4)$$

Asymptotically, multi- $\tau$  estimators converge to the true rate exponentially in  $\tau$  instead of the slow  $\tau^{-1}$ -convergence of single- $\tau$  estimators (compare to Eq. (5.5.2)). Note, that the systematic error of a multi- $\tau$  estimate thus decays much faster in  $\tau$  than its statistical error that decays by  $\tau^{-1}$  [142]. Moreover, the convergence rate of Eq. (5.5.3) is given by the gap  $\kappa_3 - \kappa_2$ . Thus, multi- $\tau$  estimators as well as single- $\tau$ -estimators are better the larger the time scale separation between the slowest and the other relaxation rates in the system is.

### 5.5.3 Existing 2-State Estimators

#### MSM based estimators

MSMs can be understood as a way of implicitly performing rate estimates via discretizing state space into small substates. Let us consider that a MSM is obtained by finely discretizing the observed space  $Y$  into crisp subsets  $Y_i \subset Y$  and estimating a transition matrix  $\mathbf{T}(\tau)$  amongst these subsets as laid out in chapter 2. It can then be shown [143] that the second eigenvector

$$\mathbf{T}(\tau)\mathbf{c} = \lambda_2\mathbf{c}$$

of this transition matrix solves the optimization problem in Eq. (5.4.5) and thus provides the best possible reaction coordinate/projection,  $\xi(y)$ , under any given observable  $y$  that are constant on each subset  $Y_i$ .

It has been suggested [53] to use the second implied time scale

$$\hat{t}_2 = -\tau / \ln(\hat{\lambda}_2(\tau))$$

as an estimate for the system's slowest relaxation time scale which correspond to the inverse relaxation rate  $\hat{\kappa}_2$  (see section 2.2). Therefore the MSM rate estimate is described by Eq. (5.5.1) with the choice  $\zeta \approx \psi_2$ . A sufficiently fine MSM thus serves an optimal single- $\tau$  rate estimator as it uses the maximum RCQ  $\hat{\alpha}$  for observables that are being discretized. However, when these observables have a poor RCQ  $\hat{\alpha}$  since they are poorly separating the slowly-converting states, there is a substantial rate estimator error according to Eq. (5.5.2) that decays slowly with  $\tau^{-1}$ . This explains the slow convergence of implied time scales shown in recent MSM studies [53, 62, 52, 51, 1] (see Figure 5.2 for an example).

#### *Estimators using dividing surfaces*

Most rate theories operate by defining a single dividing surface on the observable space  $Y$  which splits the state space into reactant A and product B. Calling

$$h_A(y) = \begin{cases} 1 & \text{if } y \in A \subset Y \\ 0 & \text{if } y \in B \subset Y \end{cases}$$

the indicator function which is 1 for set A and 0 for set B, one may define the normalized fluctuation autocorrelation function of state A [144]

$$\langle \tilde{\psi}_2(y_t) \tilde{\psi}_2(y_{t+\tau}) \rangle = \frac{\langle h_A(y_t) h_A(y_{t+\tau}) \rangle - \langle h_A(y_t) \rangle^2}{\langle h_A(y_t) \rangle - \langle h_A(y_t) \rangle^2} \quad (5.5.5)$$

In the light of our theory, Eq. (5.5.5) can also be interpreted as an autocorrelation function for the special choice

$$\zeta^{\text{divide}}(y(x)) = \frac{h_A(y(x)) - \pi_A}{\sqrt{\pi_A - \pi_A^2}}$$

where

$$\pi_A = \int_{y^{-1}(A)} dx \pi(x)$$

is the stationary probability of state A. Following Eq. (5.5.1), the procedure of splitting the observed state space into two sets and estimating the transition rate from the second eigenvalue of the 2-state transition matrix [133] results in a single- $\tau$  estimate of

$$\hat{\kappa}_2^{\text{divide}} \approx \kappa_2 - \tau^{-1} \ln \frac{\langle h_A, \psi_2 \rangle - \pi_A}{1 - \pi_A} \quad (5.5.6)$$

albeit with a deteriorated RCA of

$$\alpha^{\text{divide}} = \frac{\langle h_A, \psi_2 \rangle - \pi_A}{1 - \pi_A} \leq \hat{\alpha} \quad (5.5.7)$$

This estimator is effectively a 2-state Markov model estimate (see Ref. [133] for an analysis) and the best possible dividing surface can be chosen by maximizing  $\alpha$ , or equivalently by minimizing  $\hat{\kappa}_2$  (Eq. 5.5.1). This leads to an optimal approximation of  $\psi_2$  by two constant functions and is identical to a 2-state MSM. However, it is sub-optimal compared to the general MSM estimator that uses more states and thus a finer partition of state space and the error still decays with  $\tau^{-1}$  according to Eq. (5.5.1).

#### Signal autocorrelation estimate

Another common choice is to calculate the normalized autocorrelation function of the observed signal itself

$$\hat{\lambda}_2 = \langle \zeta^{\text{signal}}(y_t) \zeta^{\text{signal}}(y_{t+\tau}) \rangle = \frac{\langle y_t y_{t+\tau} \rangle - \langle y_t \rangle^2}{\langle y_t^2 \rangle - \langle y_t \rangle^2} \quad (5.5.8)$$

that uses

$$\zeta^{\text{signal}}(y) = \frac{y - \langle y_t \rangle}{\sqrt{\langle y_t^2 \rangle - \langle y_t \rangle^2}}$$

which is the identity of  $y$ , scaled and shifted to be properly normalized as in section 5.3.1. Like the dividing surface estimate, direct evaluation of this autocorrelation function via Eq. (5.5.1) leads to a single- $\tau$  estimate of  $\kappa_2$  that converges with  $\tau^{-1}$  as

$$\hat{\kappa}_2^{\text{signal}} \approx \kappa_2 - \tau^{-1} \ln \frac{\langle y, \psi_2 \rangle - \langle y_t \rangle \langle 1, \psi_2 \rangle}{\sqrt{\langle y_t^2 \rangle - \langle y_t \rangle^2}} \quad (5.5.9)$$

$$\approx \kappa_2 - \tau^{-1} \ln \frac{\langle y, \psi_2 \rangle}{\sqrt{\langle y_t^2 \rangle - \langle y_t \rangle^2}}. \quad (5.5.10)$$

The RCQ takes the form

$$\alpha^{\text{signal}} = \frac{\langle y, \psi_2 \rangle}{\sqrt{\langle y_t^2 \rangle - \langle y_t \rangle^2}} \leq \hat{\alpha}. \quad (5.5.11)$$

which equals one if and only if  $y = \psi_2$ .

#### Exponential-fitting estimators

Exponential-fitting estimators usually employ an exponential fit to autocorrelation functions of the form in Eq. 5.4.4 or, in order to increase numerical robustness, to their integral [135]. This can be done using different choices of  $\zeta(y(x))$  such as a step function arising from

a dividing surface, or the normalized signal itself. Figure 5.2 shows the performance of the exponential-fitting estimators on different projected dynamics.

Exponential fitting estimators are multi- $\tau$ -estimators whose systematic estimation error is given by Eq. 5.5.3. This error involves the pre-factor  $(1 - \alpha)/\alpha$ , indicating that for short lag times  $\tau$  a choice  $\xi \approx \hat{\xi}$  (where  $\alpha$  is maximized) obtained from an MSM on a fine discretization of the experimental  $y$  coordinate provides the optimal result.

#### 5.5.4 Multi-state estimation

We have demonstrated and discussed throughout this thesis that MSMs have become a popular tool in the modeling of stochastic dynamics from molecular simulations [35, 59, 1, 53, 52]. While these MSMs provide an effective way to describe the dominant transitions, the analysis of correlations functions provides a way to reconstruct the time scales of process. As we have seen in the 2-state case, multi- $\tau$ -estimations that use the information from more than one lag time provide a better rate of convergence. These ideas are the basis to estimate key properties of the multi-state MSM from correlations which we will do in the following.

As it is in general not possible to construct a Markov model for the evolution of an observation  $y_t$ , the time scales in such a MSM can only be approximately equal the time scales of the original process  $X_t$ . If our goal is to describe the evolution of projections then Markov models are not the right choice. Although it is, from a mathematical point of view, desirable to have a Markov model at hand because it allows to apply a wide variety of analysis methods (see chapter 2), one can get better estimation results when discarding the requirement of the Markov property and resorting to estimations from correlation matrices.

We now try to reconstruct the exact projected eigenvectors and the correct time scales, respectively, from an (almost) arbitrary observed time series  $y_t$ . For this we need to be aware of the following:

1. We are only operating optimally at lagtimes  $\tau > \tau_{fast}$  where we choose  $\tau_{fast}$  as large as necessary to ensure that all processes, which we declare to be too fast to be of interest, have decayed. This is a strong assumption but many systems show a separation of time scales into a few slow/interesting and a continuum of fast/negligible processes [15, 16, 17, 18, 19, 20, 21]. If this is not fulfilled, the method will still work but with a reduced rate of convergence. This is similar to a small gap in multi- $\tau$  estimation in the 2-state case.
2. The projection  $\xi$  is *non-redundant* in the dominant processes, i.e. the overlap of the projected space with the subspace of

slow processes as given in  $\mathbf{Q}$  must be non-zero. It is enough to demand that the projections of the  $n$  dominant eigenvectors  $\mathbf{Q}_{[:,i]}$ ,  $i \in \{1, \dots, n\}$  can be considered linearly independent within the numerical uncertainty. Experiments are often designed to only access a single or a small subset of the dominant processes. In this case we need to reduce the space of dominant processes to the ones *visible* by the experimental observation  $y_t$ . For the remainder we assume, that the observations/experiments are able to resolve all slowest processes of interest.

3. The projection is *stationary* over time,  $\partial_t \zeta_i(y(x)) = 0$ , which is a crucial point that can be difficult to ensure in experimental setups. This time-dependent systematic error is sometimes referred to as *drift*. In this cases the method will not fail but our error estimations are not correct. If this systematic error is small enough it may be treated well enough within the statistical uncertainty.

Using a set of observations  $\zeta$  we can compute the correlation matrices  $\mathbf{C}^X(\tau)$  as defined in Eq. (5.3.3). We will now assume, that the lag time  $\tau$  is chosen large enough, so that only a finite  $m$  dominant processes  $\psi_i$ ,  $i \in \{1, \dots, m\} \simeq M$  are present and the correlation density can be well-approximated by

$$c_\tau(x, y) \approx \sum_{i \in M} \lambda_i^\tau \pi(x) \psi_i(x) \pi(y) \psi_i(y).$$

If we define a reduced matrix  $\mathbf{Q}$  of only the dominant projected eigenvectors onto the observations by

$$Q_{ij} \equiv \langle \psi_i, \zeta_j \rangle, \quad i \in \{1, \dots, m\}, \quad j \in \{1, \dots, m\}$$

and a reduced diagonal matrix of dominant time scales  $\mathbf{\Lambda}$  by

$$\Lambda_{ii}(\tau) = \lambda_i(\tau), \quad i \in \{1, \dots, m\},$$

the correlation matrices  $\mathbf{C}^X(\tau)$  takes the simple form

$$\begin{aligned} \mathbf{C}^X(\tau) &\equiv \text{cor}[\zeta(x_t) | \zeta(x_t)](\tau) \\ &\approx \sum_{k \in M} \lambda_k(\tau) \langle \zeta, \psi_k \rangle \langle \psi_k, \zeta \rangle \\ &= \mathbf{Q}^T \mathbf{\Lambda}(\tau) \mathbf{Q} \end{aligned} \tag{5.5.12}$$

and the aim is to reconstruct  $\mathbf{\Lambda}$  and eventually also  $\mathbf{Q}$  from a given set of observation correlation matrices  $\mathbf{C}^X(\tau)$ . Even if the assumption of large enough lag times  $\tau$  is not fulfilled this definition is still reasonable as an approximation if we want the properties of the  $m$  dominant processes, which we assume in analogy to Eq. 5.5.1 in the 2-state case.

Once we have found a scheme to reconstruct  $\mathbf{Q}$  and  $\mathbf{\Lambda}$  we still need to answer, how good the estimation will be in comparison to the real dominant dynamics? The important difference compared to the construction of a discrete and reduced **MSMs** from the full dynamics (see section 2.4) is, that we are not imposing a **MSMs** anymore. We will first deal with the problem of reconstruction and address ways to estimate the approximation error by neglecting influences from fast relaxation processes later.

The following theorem proofs that in the absence of statistical uncertainty and in the ideal case of only  $m$  dominant processes being present, it is possible to exactly reconstruct  $\mathbf{Q}$  and  $\mathbf{\Lambda}$  from two observations  $\mathbf{C}^X(\tau_1)$  and  $\mathbf{C}^X(\tau_2)$  at given lagtimes  $\tau_1$  and  $\tau_2$  using a Generalized Eigenvalue Problem (**GEVP**).

**Theorem 2.** *Let  $y_t := y(x_t)$  be an observation of a Markov process  $x_t$  which can be approximated by a set of  $m$  eigenvector/eigenvalue pairs  $\{\psi_i, \lambda_i\}$  for lag times  $\tau > \tau^{\min}$  and  $\xi(y)$  a set of  $m$  observations. Then, if the matrix*

$$\mathbf{Q} : Q_{ij} = \langle \psi_i, \xi_j \rangle, i, j \in \{1, \dots, m\}$$

*is invertible and the eigenvalues  $\lambda_i$  are not degenerate, the matrix  $\mathbf{Q}$  and the time scales*

$$\mathbf{\Lambda}(\tau) \equiv \text{diag}(\lambda_1(\tau), \dots, \lambda_m(\tau))$$

*of the Markov process  $x_t$  can be reconstructed from the solutions  $\mathbf{U}$  and  $\rho_i$  of the **GEVP***

$$\mathbf{C}^X(\tau_1)\mathbf{u}_i = \mathbf{C}^X(\tau_2)\rho_i\mathbf{u}_i$$

*with any two correlation matrices,  $\mathbf{C}^X(\tau_1), \mathbf{C}^X(\tau_2) \in \mathbb{R}^{m \times m}$  at two different lagtimes  $\tau^{\min} < \tau_1 < \tau_2$ . The solutions are related by*

$$\lambda_i = \rho_i^{(\tau_1 - \tau_2)^{-1}}$$

*and*

$$\mathbf{Q} = \mathbf{U}^{-1}$$

*with the matrix  $\mathbf{U} \equiv \{\mathbf{u}_1, \dots, \mathbf{u}_m\}$  and normalized that*

$$\mathbf{u}_i \mathbf{C}^X(\tau_2) \mathbf{u}_j = \delta_{ij} \lambda_i(\tau_2)$$

*holds.*

*Proof.* First, we note, that  $\mathbf{C}^X(\tau_1)$  and  $\mathbf{C}^X(\tau_2)$  are real and symmetric, thus hermitian, and also positive definite, which follows directly from Eq. (5.5.12). Now, the **GEVP** condition

$$\mathbf{C}^X(\tau_1)\mathbf{U} - \mathbf{C}^X(\tau_2)\mathbf{U}\text{diag}(\rho_1, \dots, \rho_m)$$

is rewritten using the explicit form for  $\mathbf{C}^X(\tau)$  in Eq. (5.5.12) which yields

$$\mathbf{Q}^T \mathbf{\Lambda}(\tau_1) \mathbf{Q} \mathbf{U} - \mathbf{Q}^T \mathbf{\Lambda}(\tau_2) \mathbf{Q} \mathbf{U} \text{diag}(\rho_1, \dots, \rho_m).$$

Any non-singular matrix, i.e. also  $\mathbf{Q}$ , can be used as a congruence transformation (in the real-valued case) and using the replacement

$$\mathbf{\Lambda} \rightarrow \mathbf{Q}^T \mathbf{\Lambda} \mathbf{Q}$$

yields the simplified [GEVP](#)

$$\mathbf{\Lambda}(\tau_1) \mathbf{V} - \mathbf{\Lambda}(\tau_2) \mathbf{V} \text{diag}(\rho_1, \dots, \rho_m)$$

with a new set of generalized eigenvector  $\mathbf{V}$  or just

$$\text{diag}(\rho_1, \dots, \rho_m) = \mathbf{V}^{-1} \mathbf{\Lambda}(\tau_1 - \tau_2) \mathbf{V}. \quad (5.5.13)$$

From hermitian and positive-definite follows [\[145\]](#) that both [GEVP](#) share the same generalized eigenvalues and that the eigenvectors of the simplified [GEVP](#)  $\mathbf{V}$  are related to the original ones by

$$\mathbf{U} = \mathbf{Q}^{-1} \mathbf{V}.$$

In this simplified case the eigenvalues are simply given by

$$\rho_i = \lambda_i^{(\tau_2 - \tau_1)}$$

which can be solved for  $\lambda_i$  with

$$\lambda_i = \rho_i^{1/(\tau_1 - \tau_2)}$$

which proves the first part of the theorem. A more constructive approach is to use that the  $m$  roots of the polynomial

$$\det(\mathbf{C}(\tau_1) - \rho \mathbf{C}(\tau_2)) = 0$$

are exactly the generalized eigenvalues  $\rho_i$  [\[145\]](#). Using the congruence transformation and the explicit form of  $\mathbf{\Lambda}$  we find the same solution as given above.

To retrieve the matrix of observed eigenfunctions  $\mathbf{Q}$  we first show that the proposed normalization

$$\mathbf{u}_i \mathbf{C}^X(\tau_2) \mathbf{u}_j = \delta_{ij} \lambda_i(\tau_2)$$

can always be achieved since (1) for the generalized eigendecomposition of symmetric (hermitian) matrices and non degenerate generalized eigenvalues [\[145\]](#) follows that

$$\mathbf{u}_i \mathbf{C}^X(\tau_2) \mathbf{u}_j = 0, \quad i \neq j$$

holds and (2) an arbitrary eigenvector  $\tilde{\mathbf{u}}_i$  can be normalized by

$$\mathbf{u}_i = \tilde{\mathbf{u}}_i \sqrt{\lambda_i(\tau_2) / \tilde{\mathbf{u}}_i^T \mathbf{C}^X(\tau_2) \tilde{\mathbf{u}}_i}.$$

From the simplified [GEVP](#) in Eq. [\(5.5.13\)](#) we directly find the matrix of eigenvectors  $\mathbf{V} = \text{Id}$  to be the identity matrix, which leads to a

possible matrix of eigenvectors given by  $\hat{\mathbf{U}} = \mathbf{Q}^{-1}\text{Id} = \mathbf{Q}^{-1}$ , which is unique up to a normalization if the eigenvalues are non-degenerate. The normalization at lag time  $\tau_2$  immediately ensures that the choice of  $\mathbf{U}$  is unique and that

$$\mathbf{U}^T \mathbf{C}^X(\tau_2) \mathbf{U} = \mathbf{\Lambda}(\tau_2)$$

holds, which is equivalent to

$$\mathbf{C}^X(\tau_2) = \mathbf{U}^{-T} \mathbf{\Lambda}(\tau_2) \mathbf{U}^{-1}. \quad (5.5.14)$$

Using the Generalized Eigenvalue Problem (GEVP) formulation

$$\begin{aligned} \mathbf{C}^X(\tau_1) \mathbf{U} &= \mathbf{C}^X(\tau_2) \mathbf{U} \text{diag}(\rho_1, \dots, \rho_m) \\ &= \mathbf{C}^X(\tau_2) \mathbf{U} \mathbf{\Lambda}(\tau_1 - \tau_2) \end{aligned}$$

in Eq. (5.5.14) we derive

$$\begin{aligned} \mathbf{C}^X(\tau_1) &= \mathbf{U}^{-T} \mathbf{\Lambda}(\tau_2) \mathbf{U}^{-1} \mathbf{U} \mathbf{\Lambda}(\tau_1 - \tau_2) \mathbf{U}^{-1} \\ &= \mathbf{U}^{-T} \mathbf{\Lambda}(\tau_2) \mathbf{\Lambda}(\tau_1 - \tau_2) \mathbf{U}^{-1} \\ &= \mathbf{U}^{-T} \mathbf{\Lambda}(\tau_1) \mathbf{U}^{-1} \end{aligned}$$

so that the uniquely determined  $\mathbf{Q} = \mathbf{U}^{-1}$  and  $\mathbf{\Lambda}$  under the given normalization will correctly reconstruct  $\mathbf{C}^X(\tau_1)$  and  $\mathbf{C}^X(\tau_2)$ . Henceforth  $\mathbf{Q}$  and  $\mathbf{\Lambda}$  must be the correct unique parametrization of the initial Markov process  $X_t$ .  $\square$

With this theorem we have shown that under the given simplifications, especially of large enough lag times  $\tau$  and absence of statistical uncertainty, we can reconstruct the dominant time scales and the projections of eigenvectors under almost any set of observations. This idea will be the basis for the multi-state multi-lag time estimation which we call *spectral estimation*.

### 5.5.5 Optimizational Approach

Proven, that the estimation works for an arbitrary choice of two lag-times, still, in the presence of statistical uncertainty the results of these estimations will not agree. Instead of relying on the minimal requirement of data to fit the model parameters, the problem can also be stated as an optimization problem where more than only two different correlation matrices  $\mathbf{C}^X(\tau)$  are used to find the parameters  $\mathbf{Q}$  and  $\mathbf{\Lambda}$

$$\{\hat{\mathbf{Q}}, \hat{\mathbf{\Lambda}}\} = \underset{\mathbf{Q}, \mathbf{\Lambda}}{\text{argmin}} \sum_{\tau \in \{\tau_{\min}, \tau_{\max}\}} \left\| \mathbf{Q}^T \mathbf{\Lambda}(\tau) \mathbf{Q} - \mathbf{C}^X(\tau) \right\|.$$

This is the extension of multi- $\tau$ -estimation in the 2-state case in section 5.5 to more than two states.

Recalling that different dominant processes are observable at different time scale regions (see Figure 2.3) we need to be concerned that the number of present processes should not change in the optimization if we want all processes to be treated equally important. If e.g. the range of lag times is extended into a region, where only the stationary process is present, than fitting will be dominated by the stationary process. To compensate for this problem an iterative scheme might be a solution: Use the optimization on a large range of lag times, and estimate all processes and keep only the dominant one. Repeat the optimization on a smaller time scale range and keep the already estimated process vectors and time scales fixed. This way the maximal range for the estimation of each process can be used provided that the time scale ranges are chosen carefully! The optimization approach has to be investigated further and is subject of current research.

## 5.6 SENSITIVITY ANALYSIS

In real world applications we need to take into account that we base our model on finite data. This means, that the observations we use to estimate the time scales from can only be approximately correct, even if an experimental setup would be able to measure without error. This problem might be serious since we have not yet excluded the possibility that even very small deviations in the observed correlations might lead to large errors in the estimated variables. We can address the problem of uncontrolled errors by using the fact that both  $\mathbf{Q}$  and  $\mathbf{\Lambda}$  are continuous functions of the two observations  $\mathbf{C}(\tau_1)$  and  $\mathbf{C}(\tau_2)$  so that given a specific error bound for the estimated variables there exists also a bound on the correlation functions to ensure this quality. We proceed with an sensitivity analysis as we have done before in chapter 3 for the committor. To analyze the asymptotic behavior of the Generalized Eigenvalue Problem (GEVP) we proceed with the derivation as outlined in [146]:

**Lemma 3. Generalized Eigenvalue Sensitivity** *If the two (symmetric) observation correlation matrices  $\mathbf{A}(\theta) := \mathbf{C}(\tau_1; \theta)$  and  $\mathbf{B}(\theta) := \mathbf{C}(\tau_2; \theta)$  are depending on a scalar parameter  $\theta$  and  $\rho$  and  $\mathbf{y}$  solve the Generalized Eigenvalue Problem (GEVP)*

$$\begin{aligned} (\mathbf{A} - \rho\mathbf{B})\mathbf{y} &= 0 \\ \mathbf{y}^T\mathbf{B}\mathbf{y} &= 1 \end{aligned}$$

then the linear sensitivity  $\partial_{\theta}\rho$  is related to  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\lambda$  and  $\mathbf{y}$  by

$$\frac{\partial\rho}{\partial\theta} = \mathbf{y}^T \left( \frac{\partial\mathbf{A}}{\partial\theta} - \rho \frac{\partial\mathbf{B}}{\partial\theta} \right) \mathbf{y}.$$

If  $\mathbf{A}$  or  $\mathbf{B}$  depend on more than one parameter these will be computed independently and then be added up. The case that  $\mathbf{A}$  and  $\mathbf{B}$  depend on different variables can be treated in the same way.

We can in addition prove a similar lemma for the process vectors  $q_i$  which are related to the eigenvectors by a simple inversion. We write a simple lemma and proof

**Lemma 4. Process Vector Sensitivity** *If the two (symmetric) observation correlation matrices  $\mathbf{A}(\theta) := \mathbf{C}(\tau_1; \theta)$  and  $\mathbf{B}(\theta) := \mathbf{C}(\tau_2; \theta)$  are depending on a scalar parameter  $\theta$  and  $\rho$  and  $\mathbf{y}$  solve the Generalized Eigenvalue Problem (GEVP)*

$$\begin{aligned} (\mathbf{A} - \rho\mathbf{B})\mathbf{y} &= 0 \\ \mathbf{y}^T\mathbf{B}\mathbf{y} &= 1 \end{aligned}$$

or more generally for the full set of generalized eigenvalues given by the diagonal matrix of generalized eigenvalues  $\mathbf{\Lambda}$  and the matrix of column-wise generalized eigenvectors  $\mathbf{Y}$

$$\begin{aligned} \mathbf{Y}^T(\mathbf{A} - \mathbf{\Lambda}\mathbf{B})\mathbf{Y} &= 0 \\ \mathbf{Y}^T\mathbf{B}\mathbf{Y} &= Id \end{aligned}$$

then the generalized eigenvector sensitivity  $\partial_\theta\mathbf{y}$  is related to  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\rho$  and  $\mathbf{y}$  by

$$\frac{\partial\mathbf{y}}{\partial\theta} = -(\mathbf{A} - \rho\mathbf{B})^{-1} \left( \frac{\partial\mathbf{A}}{\partial\theta} - \rho \frac{\partial\mathbf{B}}{\partial\theta} \right) \mathbf{y} - \frac{1}{2} \left( \mathbf{y}^T \frac{\partial\mathbf{B}}{\partial\theta} \mathbf{y} \right) \mathbf{y}.$$

where

$$(\mathbf{A} - \rho\mathbf{B})^{-1} := \mathbf{Y}[\mathbf{\Lambda} - \rho Id]^+ \mathbf{Y}^T$$

is the generalized symmetric inverse with  $[\cdot]^+$  representing the Moore-Penrose inverse which can in this case be written as

$$[\mathbf{\Lambda} - \rho Id]^+ = \text{diag} \left( \begin{cases} \rho_i \neq \rho & \rho^{-1} \\ \text{else} & 0 \end{cases} \right).$$

The proofs can be found in appendix C.1.

Using Lemma 3 and the results from Theorem 2 we find for the relative sensitivity of estimated eigenvalues that

$$\frac{\partial_\theta\lambda}{\lambda} = (\tau_2 - \tau_1)^{-1} \mathbf{y}^T \left( \rho^{-1} \frac{\partial\mathbf{C}^X(\tau_2)}{\partial\theta} - \frac{\partial\mathbf{C}^X(\tau_1)}{\partial\theta} \right) \mathbf{y}$$

holds which implies that the statistical sensitivity decreases with larger lag time differences.

## 5.7 SPECTRAL ESTIMATION

Having laid out a theoretical basis we will now propose a scheme to apply the new spectral estimation to an observed (experimental) time series  $y_t$ .

1. Decide on a number of processes  $n$  you wish to analyze (analyze spectrum, etc.). This can be done by using the ideas from the spectral decomposition of a Markov model (see chapter 2) or by experience.
2. Discretize the observed time series/trajectory  $y_t \in \mathbb{R}$  into a larger number  $m \geq n$  of micro bins (e.g. 100). This discretization will induce the basis  $\mathcal{G}(y)$  for functions that are constant on each of the  $m$  micro bins.
3. Construct a Markov model using the correlation matrix  $\mathbf{C}^G(\tau)$  which is derived from the projection onto the basis  $\mathcal{G}(y)$  by

$$\mathbf{T}^G \equiv (\mathbf{\Pi}^G)^{-1} \mathbf{C}^G(\tau)$$

with  $\mathbf{\Pi}^G = \text{diag}(\pi_1^G, \dots, \pi_n^G)$  being the diagonal matrix of the projected stationary distribution.

4. Compute the matrix of the  $n$  dominant right eigenvectors  $\mathbf{A} \in \mathbb{R}^{m \times n}$  (in columns) and eigenvalues  $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$  defined by

$$\mathbf{T}^G \mathbf{A} = \mathbf{A} \mathbf{\Lambda}^G$$

at some lag time  $\tau$  where all relevant processes are assumed to be observable. The construction of a reversible transition matrix  $\mathbf{T}^G$  assures, that the right eigenvectors of  $\mathbf{T}^G$  will be orthogonal w.r.t. the observed stationary distribution  $\pi^G$  of the Markov model. Note, that up to now this is the same procedure used to compute Implied Time Scale (ITS)

$$t_i = -1 / \log(\lambda_i^G).$$

5. Using a projection induced by the right eigenvectors

$$\xi_i = \sum_{j=1}^m A_{ji} g_j(y), \quad n \in \{1, \dots, n\}$$

compute the correlation matrices

$$\begin{aligned} \mathbf{C}_{ij}^X(\tau) &= \text{cor}(\xi_i(y_t), \xi_j(y_{t+\tau})) \\ &= \mathbf{A}^T \mathbf{C}^G(\tau) \mathbf{A} \end{aligned}$$

for a number of lagtimes  $\tau \in L$ ,  $\tau_{\min} < \tau < \tau_{\max}$ .

6. Optimize the matrix of projected eigenvectors  $\mathbf{Q}$  and the matrix of eigenvalues

$$\Lambda(\tau) \equiv \text{diag}(\lambda_1^\tau = 1, \lambda_2^\tau, \dots, \lambda_n^\tau)$$

to minimize

$$\{\hat{\mathbf{Q}}, \hat{\Lambda}\} = \underset{\mathbf{Q}, \Lambda}{\text{argmin}} \sum_{\tau \in \mathcal{L}} \left\| \mathbf{Q}^T \Lambda(\tau) \mathbf{Q} - \mathbf{C}^X(\tau) \right\| \quad (5.7.1)$$

to get the final approximation of the original eigenvalues of the hidden process  $X_t$ .

Alternatively, in the case of only two lagtimes  $\tau_1$  and  $\tau_2$ , the problem can be solved directly using a [GEVP](#) solver for hermitian matrices and the normalization from [Theorem 2](#) to compute  $\mathbf{Q}$  and  $\Lambda$ .

7. Repeat from step 5 with variations in the range  $\{\tau_{\min}, \tau_{\max}\}$  of lagtimes used or from step 2 which different numbers of dominant processes  $n$ .

For a simplified graphical illustration see left side of [Figure 5.9](#).

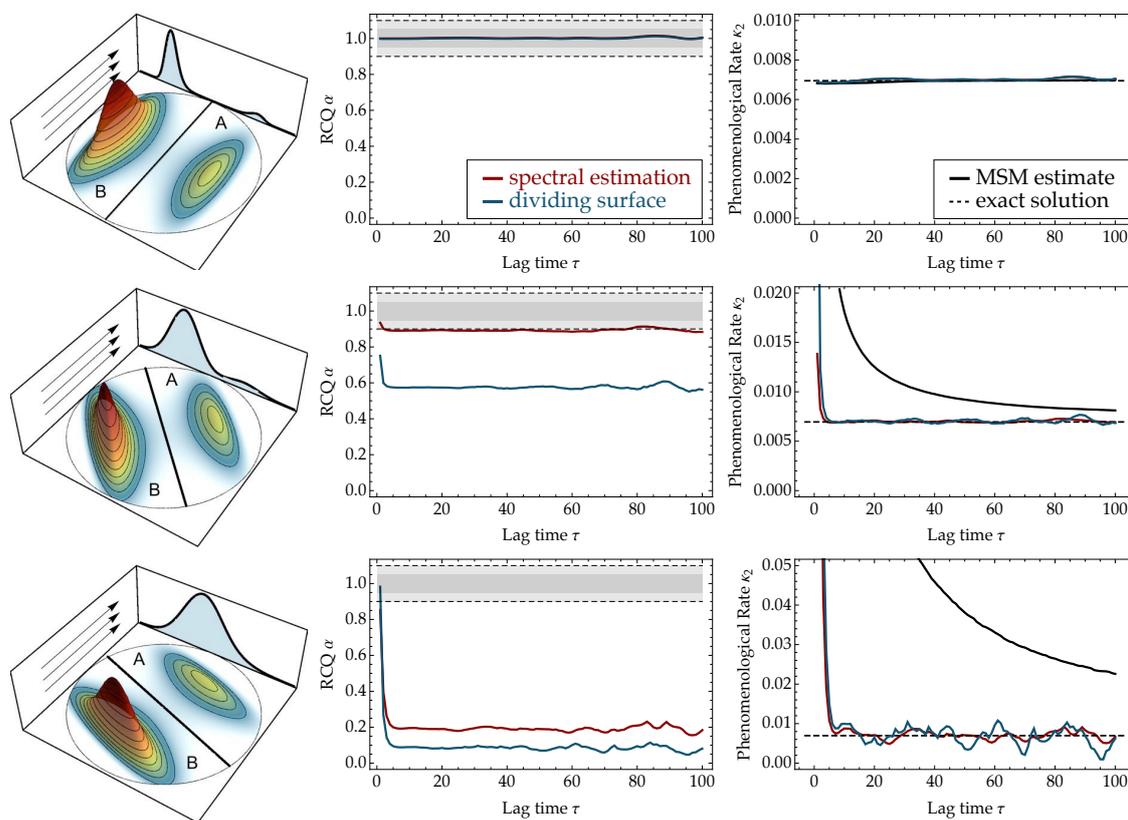
## 5.8 EXAMPLES

### 5.8.1 2-state example in 2D

For a demonstration of the improved convergence of multi- $\tau$ -estimators in the in the 2-state case we computed a 10,000,000 step test trajectory in a simple 2D energy landscape for a diffusional process (details can be found in [appendix D](#)). [Figure 5.2](#) shows the estimation results for two different multi- $\tau$ -estimators and a [MSM](#)-based single- $\tau$ -estimators that effectively computes the slowest [ITS](#). Depending on the projection angle (left column) the [MSM](#) based estimation converges much slower than the spectral estimation. Even in the case where both states almost totally overlap the multi- $\tau$ -estimators provide a quite good estimation, although the susceptibility to statistical fluctuations increases. Finally,  $\hat{\alpha}$  can also be determined by the spectral estimation procedure in the 2-state case. [Figure 5.2](#) shows estimates of the [RCQ](#)  $\alpha$  of different projected dynamics using different rate estimators and we recover the fact that the projection onto an approximate eigenvector (red) provides better [RCQ](#) than using a dividing surface (blue).

### 5.8.2 4-state example in 1D

For a proof of concept of the multi-state case we use the same 4-state example previously introduced in [chapter 2](#) (see [Figure 2.1](#) and [appendix D](#) for details) and try to estimate the correct time scales from



**Figure 5.2 – 2-state rate estimation example.** Comparison of the performance in the 2-state case for a 10,000,000 step test trajectory. Columns: Angle of projection of 2D dynamics. For details see appendix D). The greater the angle, the more difficult it is to separate both metastable states in the observation. (left column) projection of 2D dynamics, (middle column) estimated reaction coordinate quality  $\alpha$ , (right column) time scale estimation. (red) Estimation from Projection onto 2nd eigenvector. (blue) Estimation from projection using a dividing surface. (black) Estimation from 2nd eigenvalue of the transition matrix estimated the given lagtime  $\tau$ . Even in the case where both states almost totally overlap the rate estimation is quite good, while the implied time scale Implied Time Scale approach fails.

a set of 5 different projections onto 4-states. We assume no statistical error which is equivalent to the assumption that the used observation correlation matrices were estimated from an infinitely long trajectory. Our estimation is solely based on the projected correlation matrices  $C^X(\tau)$  given at various lagtimes  $\tau$ . Figure 5.3 shows the results for all 5 projections comparing the estimations using the new method and the implied time scales obtained from a MSM constructed from the projected dynamics.

The first row presents an illustration of the projection with the four grey shaded functions corresponding to the four projection functions  $\zeta_i$  while the colored regions are the projection of the stationary distribution onto these four states. This indicates where in the energy landscape the projections are located, how reasonable they correspond to the metastable regions and how well they are separated.

The second row presents the 4 dominant eigenfunctions projected onto the observables and the colored regions depict the projection error, which indicate that especially the good crisp clustering (col 1) and the PCCA projection (col 3) are good choices in reducing the projection error.

Row three shows the *presence*  $\gamma_i$  of a process which we define as

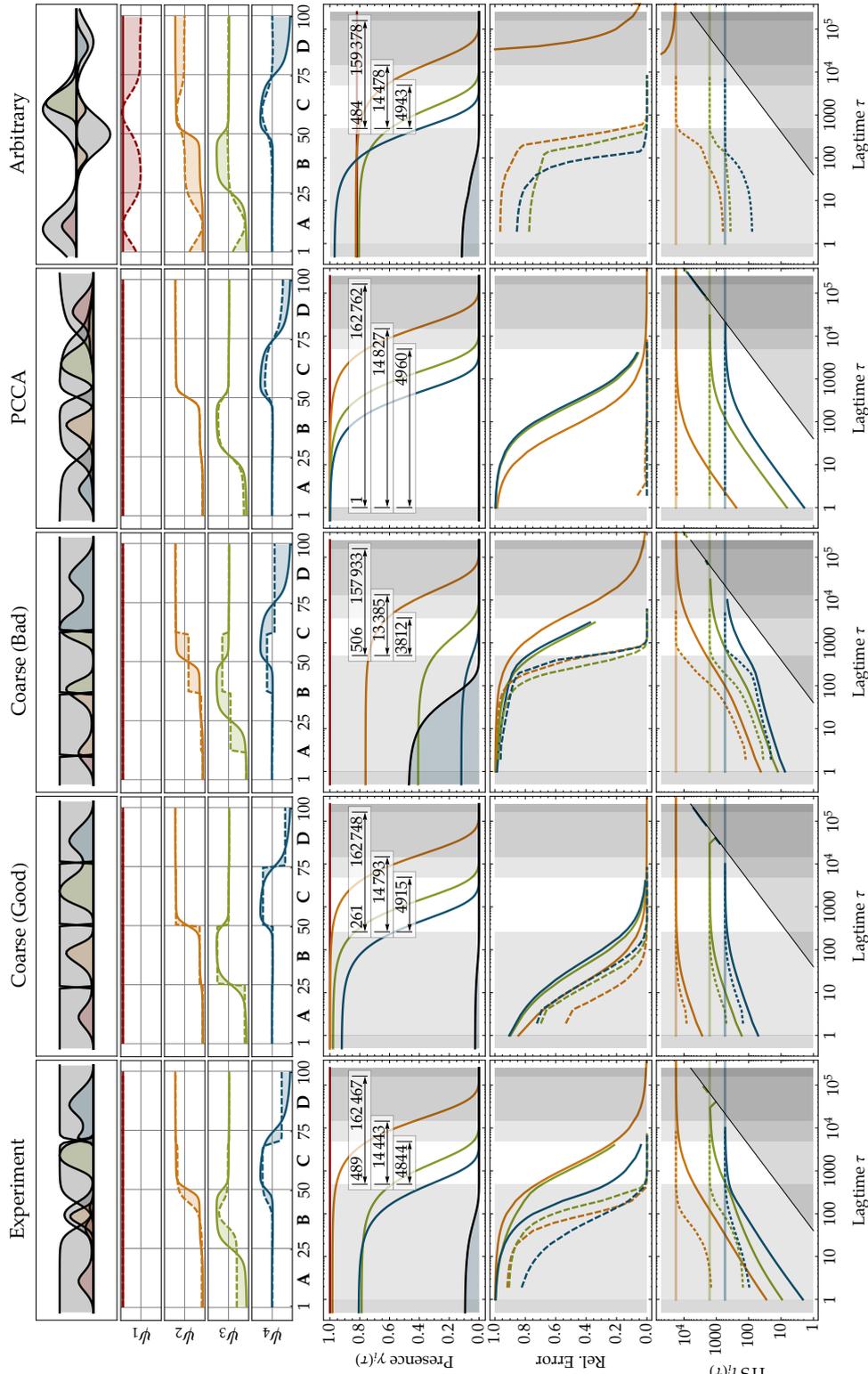
$$\begin{aligned}\gamma_i(\tau) &= \lambda_i(\tau) \cdot \|X\psi_i\|^2 \\ &= \lambda_i(\tau) \cdot \|(\text{Id} - X^\perp)\psi_i\|^2 \\ &= \lambda_i(\tau) \cdot (1 - \delta_i^2),\end{aligned}$$

the squared length of the  $i$ -th eigenvector projected by  $X$  in Eq. ((2.4.3)) into the space spanned by all observables weighted with the eigenvalue  $\lambda(\tau)$  at lag time  $\tau$ . This value is a measure of how much a certain process (the eigenvector/eigenvalue pair) is present in the observation at a certain lagtime. If  $\gamma_i(\tau)$  is small than the process has disappeared and cannot be reconstructed from the observation. The presence  $\gamma_i$  is directly related to the projection error  $\delta_i$  defined in Eq. (2.4.5). This is consistent with the two possibilities to reduce the appearance of a certain process, either increase the lagtime or choose a projection orthogonal to the relevant eigenvector  $\psi_i(x)$ .

In the case of a membership projection (defined in Eq. (5.3.5)), i.e. a projection that contains the constant projection  $\mathbf{1}$ , the projection error for the stationary distribution is zero,  $\delta_1 = 0$ , and it can be shown that  $\gamma_1 = \lambda_1(\tau) = 1$ . The relevance plots (row 3) can hence be used to displays the range of time scales where a process is present  $\gamma_i(\tau) > \gamma^{\min}$  in the observation. The left grey-shaded region refers to the time scales, where

$$\max_{i>4} \gamma_i > \gamma^{\min} = 10^{-4}$$

and thus any of the considered fast processes is present. Respectively, the right gray-shaded regions refer to the parts where not all dominant processes are present  $\gamma_i < \gamma^{\min}$ ,  $i \leq 4$ . Depending on the projection the range extends in both directions, especially the eigenvector



**Figure 5.3 – Time scale Estimation** Comparison of the implied time scales estimated using spectral estimation and using a construction of a Markov model. (columns) Various projections, (rows) [1st] illustration of projection, [2nd] projected eigenvectors, [3rd] Relevance  $\gamma_i$  indicating the presence of processes in the projection vs. lagtime  $\tau$ , [4th/5th] Relative Error and Absolute estimated time scales  $t_i$ . Constant lines indicate the exact solution. Colors indicate the eigenvector/process considered 1/red, 2/yellow, 3/green, 4/blue. The spectral estimation (dashed lines) performs much better than the implied time scale (ITS) estimates (solid lines). As soon as the presence of fast processes vanishes the spectral estimation becomes exact. The gray shaded regions refer to time scale ranges where certain processes can be observed.

based projections (PCCA in column 3) has no fast processes present and is thus perfect even for small time scales.

Finally the estimated time scales  $t_i$  (row 5) and their relative errors (row 4) are presented. The dashed lines refer to the spectral estimation, while the continuous lines are estimated using implied time scales from a projected MSM. The results are shown versus an implied time scale. Since spectral estimation uses two lagtimes, so that the estimated value is given at the higher lag time  $\tau_2$  while the first lag time is chosen to be  $\tau_1 = \tau_2/2$ . In principle one could obtain even better results when increasing  $\tau_1$  towards  $\tau_2$ . Nevertheless, the method provides excellent improvements over the MSM approach. Close to time scales where the fast processes have vanished, the error reduces almost to zero. This is in agreement with the results from Theorem 2. Also, the time scales are never overestimated which happens for non-membership projections in the MSM case, since the stationary process mixes with the other processes. Lastly and most importantly the method provides always better estimations even for small lag times compared to the MSM-based estimation. As soon as the fast processes have decayed (white region) the estimation error almost vanishes. The diagonal crossing gray region refers to the time scales not accessible due to numerical machine precision of the used eigenvalue solver. Assuming that the smallest non-zero computed eigenvalue (or the root of the polynomial) is  $\epsilon$  then the smallest possible implied time scale computation at a discrete lagtime multiple of  $k$  will lead to

$$t^{\min}(k) = -\frac{k}{\log(\epsilon)}$$

which gives with a precision of  $\epsilon \approx 3 \cdot 10^{-18}$  roughly

$$t^{\min}(k) = 0.0249 k$$

and explains the linear cut-off behavior observed in some cases for large  $\tau$  estimations.

## 5.9 APPLICATION TO OPTICAL TWEEZER EXPERIMENTS

### 5.9.1 *Experimental Setup*

As an experimental proof of concept we are analyzing a set of time series that have been recorded using a constant trap position (equilibrium) optical tweezer experiment of a 56-base ssRNA hairpin, the schematic set up of which is shown in Figure [5.4] (see appendix D.2 and [147] for details).

This experiment provides us with a set of trajectories that measure the force exerted on an attached bead inside a laser trap with approximately harmonic potential and thus indirectly track the end-to-end distance of the hairpin. Thus the data collected is a projection of the

internal full dynamics of the ssRNA hairpin onto a one-dimensional observable. which allows to us to apply the new method of spectral estimation to predict time scales of relaxation processes that affect this specific observable. We expect that these will be the processes that affect the end-to-end distance of the probed molecule.

### 5.9.2 Time scale Estimation

The estimation procedure described in section 5.7 was applied to a set of six time series which were recorded in an optical tweezer experiment from a single hairpin. Each time series was observed at room temperature ( $T = 293\text{ K}$ ) and in equilibrium (after an equilibration phase) from a single unique RNA fiber with only various changed total trap extension. Refer to appendix D.2 and the corresponding publication [147] for details about the experimental setup. The six resulting time series, as displayed in Figure 5.5, were sampled for 60s at a time resolution of 0.020 ms (corresponding to a sampling frequency of 50 kHz) giving a series of  $3 \cdot 10^6$  time points in a range between 120 nm and 165 nm for the relative bead displacement. Actually recorded were forces which have then been converted to bead displacements relative to the trap center using Eq. 5.10.1. Each time series was binned into a set of 50 equally spaced states over the full range of observed data points in the particular time series to achieve a comparable spatial resolution. Afterwards transition matrices were estimated at the native sampling lagtime of  $\tau_0 = 0.02\text{ ms}$  and the first 5 eigenvectors  $\hat{\psi}_i$  were computed to be used as a basis for the used observation  $\xi$ . All time series were projected onto this 5 observations  $\xi_1, \dots, \xi_5$ . Then, two times a set of correlation matrices  $\mathbf{C}^X(\tau)$  were computed for 20 lag times in a range of

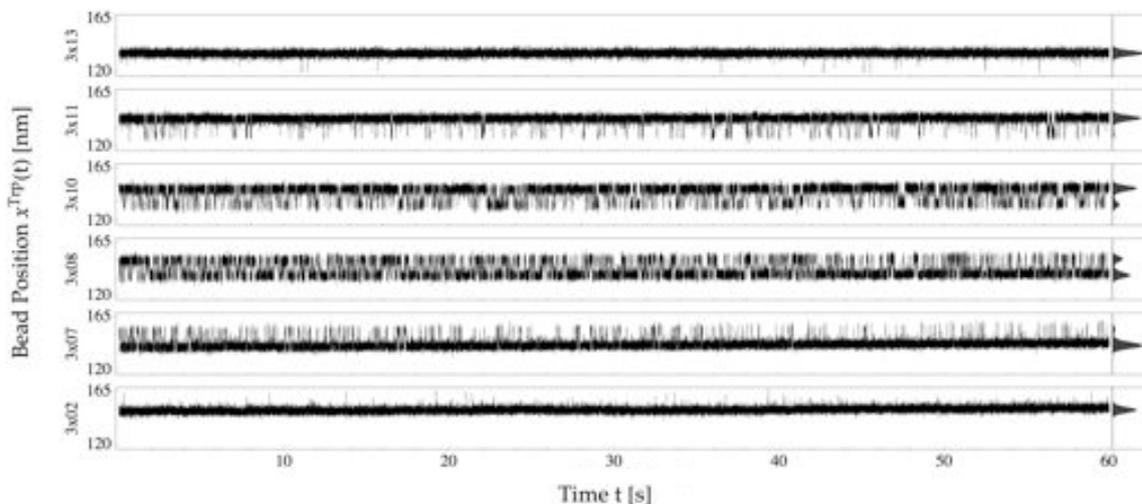
$$\tau \in \{10\tau_0, \dots, 200\tau_0\} = \{0.2\text{ms}, \dots, 4.0\text{ms}\},$$

the first one using all 5 eigenvectors,  $\xi^5 = \{\hat{\psi}_1, \dots, \hat{\psi}_5\}$ , and the second one only the 4 dominant ones,  $\xi^4 = \{\hat{\psi}_1, \dots, \hat{\psi}_4\}$ .

From these  $2 \times 20$  correlation matrices the time scales for the 6 tweezer positions were computed using the spectral estimation method in section 5.7 by the optimization of Eq. 5.7.1 to reduce statistical influences. Here, a set of various lag time ranges  $L$  was used to feed the optimization: Changing the start of the range, over which the optimization was run, between  $\tau^{\min} \in \{10\tau_0, \dots, 170\tau_0\}$  while keeping the end of the range fixed at  $\tau^{\max} = 200\tau_0$ . The step size was also varied and set to  $\Delta\tau \in \{2, 4, 5, 8, 10\} \cdot \tau_0$ . This procedure led to a total of  $2 \times 85$  optimizations in which the number of used correlation matrices varied between 4 and 96.

To execute the actual optimization of Eq. 5.7.1, the build-in minimization algorithm from Mathematica [148] (command FindMinimum) was used with the maximum number of iterations increased to 2000





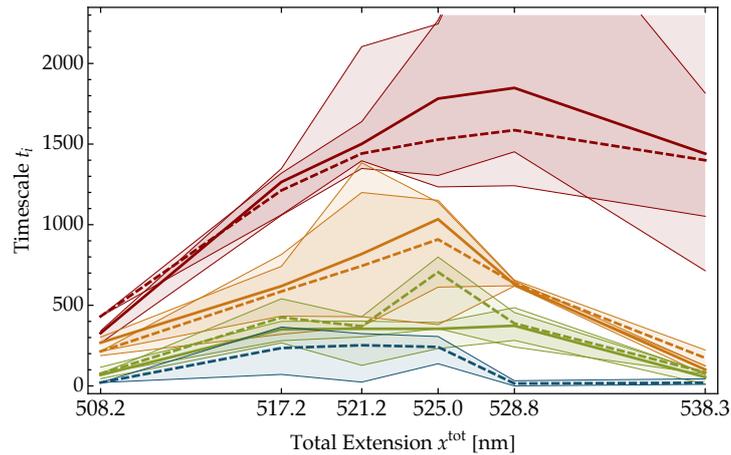
**Figure 5.5 – Experimental traces** Experimentally measured fluctuations in the bead position (relative to the trap center) computed from the traced forces in the optical tweezer experiment for 6 selected different total trap extensions. The index on the left indicates the trace in the original experiment for reference (see appendix D.2 and Ref. [147]). All traces were recorded from the same hairpin for 60 s at a sampling frequency of 50 kHz resulting in  $3 \cdot 10^6$  time points per trace. Depending on the tweezer position the stationary distribution shows a clear two state separation.

(option `MaxIterations`). The algorithm reports if it has converged within the standard accuracy, which was set to half the number of digits of the used machine precision – in the present case about  $\sim 10^{-8}$ . Only if this was the case the estimated time scales were kept for later analysis.

The results of this time scale estimation are presented in Figure 5.6, where the thick lines indicate the median over all valid estimations and the colored areas refer to the 25%/75% quantile regions. Both estimations, for four ( $\zeta^4$ ) and five ( $\zeta^5$ ) projection functions, provide a smooth estimate and agree well, although the 5-state estimation seems to provide slightly slower timescales compared to the 4-state case. A possible explanation is that to resolve the 5th slowest process the lower end  $\tau^{\min}$  of the range of lagtimes  $L$  had to be in a range where this process is still present. If  $\tau^{\min}$  is too large the optimization effectively tries to fit 5 processes where only 4 are present and does not converge. Hence, the lag time ranges  $L$  with high  $\tau^{\min}$  are excluded from the estimation and thus the systematical error induced by the presence of fast processes might be stronger in this case.

## 5.10 THEORETICAL RNA MODEL

To get a more detailed picture of the dynamics, it would be useful to associate the relaxation processes that were found using spectral estimation to actual structural rearrangement in the RNA hairpin. For this we employ a phenomenological energy model of a RNA hairpin



**Figure 5.6 – Estimated time scales vs (estimated) total trap extension** for 6 selected experimental traces. Shown are estimations for four  $\zeta^4$  (solid lines) and five  $\zeta^5$  (dashed lines) projection functions. Colors indicate the actual time scale ordered from slow to fast: slowest  $t_2$  (red),  $t_3$  (yellow),  $t_4$  (green),  $t_5$  (blue). The thick lines indicate the median over all estimations with 25%/75% quantile regions indicated by colored areas.

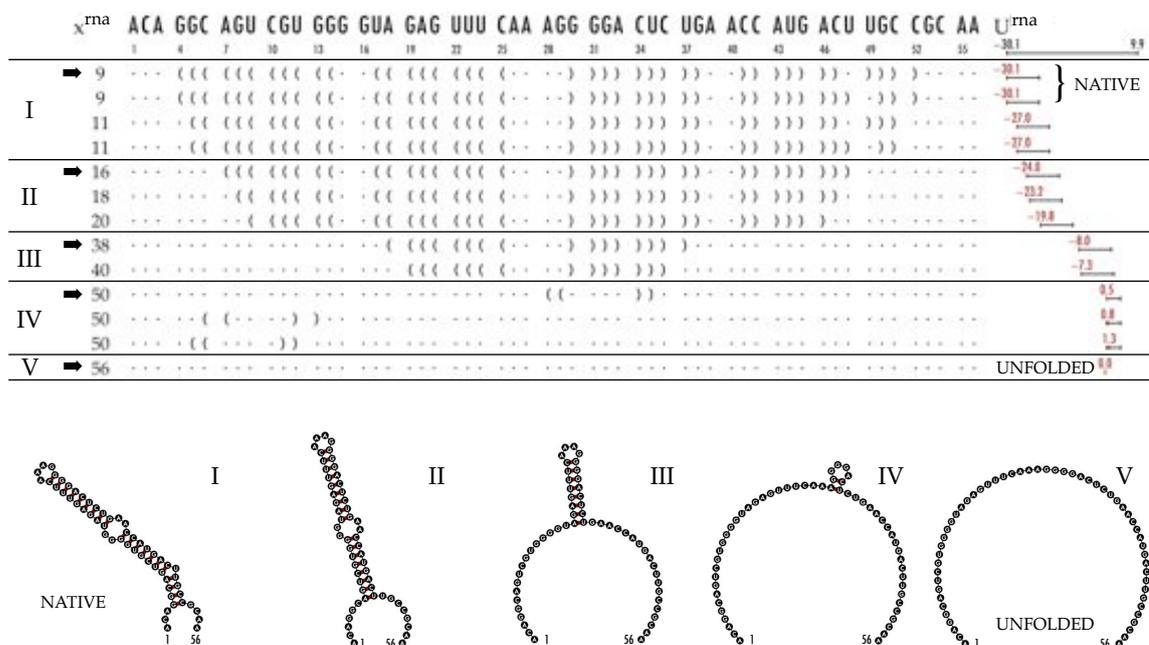
that is being subject to the force probe experiments analyzed here. In the following we will review the necessary model components.

#### 5.10.1 RNA configurations

We decided to use the stable RNA structures predicted from the Vienna RNA Web Service[149]. This project, which is publicly available as a web service, allows, using sophisticated empirical forcefields, to compute possible configurations of RNA chains with their relative potential energies. In the present case the ssRNA strain is a p21ab RNA Hairpin which can be represented by a sequence of 56 single RNA bases as given in Figure [5.7].

Various RNA configurations are possible due to a multitude of possible ways of pairing the bases by bonds (see Figure 5.7 for some examples). To limit the number of conformations, the model assumes that a configuration is solely determined by the specification of these bond pairings and, in addition, that the configuration can be drawn on a 2D plane without twists and crossings (topological order zero)[149]. This allows to represent a configuration conveniently by a sequence of characters consisting of parenthesis (indicating a bond) and dots (indicating bases without connection) (see Figure [5.7]). To retrieve the set of possible RNA configurations we used the RNAsubopt command

```
RNAsubopt -d2 --noLP -s -e 40 < seq.inp > rna.out
```



**Figure 5.7 – Classification of structures of the p21ab RNA Hairpin.** The upper table lists the structure in the form used by the Vienna RNA Web Service to express the position of bonds between RNA bases.  $x^{rna}$  denotes the free end to end distance in terms of bases relevant for the trap setting. The energy  $U^{rna}$  on the right lists the range of energies occurring in the model at the fixed end-to-end distance in kcal mol<sup>-1</sup>. The energies are normalized to be zero for the completely unfolded state (V). All configurations are manually lumped into 5 groups (I-V) depending on the end-to-end distance indicating folded (I) to unfolded (V) - see Tab. 5.2 for definition. The schematic representations at the bottom show the lowest energy configuration within each group as vertex-free planer structures which are marked with an arrow in the table above.

which was set to use a cut-off of  $40 \text{ kcal mol}^{-1}$  above the minimum free energy and the sequence of the RNA hairpin seq.inp of

```
ACA GGC AGU CGU GGG GUA GAG UUU CAA AGG GGA CUC UGA
ACC AUG ACU UGC CGC AA.
```

In our case a total of 6,087,011 configurations between  $-30.1 \text{ kcal mol}^{-1}$  and  $+9.9 \text{ kcal mol}^{-1}$  were predicted.

Assuming a canonical ensemble of configurations the relative probability to observe a specific configuration  $c$  is given by the Boltzmann-Distribution related to the energy  $U_c^{\text{rna}}$  representing the internal bonds by

$$p_c(\beta) \propto \exp(-\beta U_c^{\text{rna}})$$

where  $\beta$  is again the inverse temperature in units of the Boltzmann constant  $k_B$ . Here, the system is coupled to optically trapped polystyrene beads with double-stranded DNA (dsDNA) handles that are used to change these probabilities according to a specific trap setting.

What effectively happens is that applying a pulling force tilts the energy landscape towards configurations that have a larger end-to-end distance  $x^{\text{rna}}$ . To treat the effects of the different experiment components (e.g. trap, handles, hairpin) onto  $x^{\text{rna}}$  correctly we have calculated the energy dependency for all these components separately. To keep things comparable and avoid errors, the same units for all constants will be used; we decided for  $\text{kcal mol}^{-1}$  for energies and nm for distances. We will assume that energy is conserved and can travel freely between different components and that the energy of the system components can be described *well enough* by a single extension number  $x$ , thus, assuming that all parts are aligned in a straight line.

### 5.10.2 The optical trap

The trap is assumed to be harmonic (i.e. linear increase of the force with increasing bead displacement in the trap) which allows us to compute the displacement of the center of the optically trapped bead from the trap center from the measured force on the bead,  $x^{\text{trp}}$ , once the trap spring constant has been determined. In the following sections we will exclusively use distances/displacements instead of the actual measured forces. This approximation is only valid within a certain range of forces which are met in this case. Exceeding this range, the force of the trap reaches a saturation region and the forces remain almost constant with increasing displacement  $x^{\text{trp}}$  from the equilibrium. Since the forces exerted here are small enough for the system to remain in the harmonic region we model the energy of the trap by

$$U^{\text{trp}}(x^{\text{trp}}) = 1/2 k^{\text{Trap}} (x^{\text{trp}})^2 \quad (5.10.1)$$

where we assume a force constant of

$$k^{\text{Trap}} = 0.096 \text{ pN/nm}$$

which equals

$$k^{\text{Trap}} = 0.014 \text{ kcal mol}^{-1} / (\text{nm})^2.$$

The force constant was initially set to a value given from the experiment and later slightly optimized to match the observed traces best (see appendix D.2 for details).

### 5.10.3 The dsDNA Handles

The handles resemble a much more complicated behavior and can be modeled by a polymer chain model [150]. An approximation of the force needed to pull on such a polymer chain in the case of high forces is given by a non-closed form

$$\beta L_p \cdot F(x^{\text{hdl}}) = \frac{1}{4(1 - x^{\text{hdl}}/L_0 + F(x^{\text{hdl}})/E_y)^2 - 1/4 + x^{\text{hdl}}/L_0 - F(x^{\text{hdl}})/E_y}$$

with additional bond-stretching corrections. Here  $L_p = 45 \text{ nm}$  refers to a typical persistence length of the dsDNA chain [151] and  $L_0$  is the contour length, the total length of the backbone. In the present case, handles of 558 base pairs are used with a typical length per base pair of 0.34 nm, leading to a contour length of

$$L_0 = 189.7 \text{ nm}$$

per DNA chain and finally, for the Young modulus  $E_y$  for phosphodiester bond stretching a typical value of

$$E_y = 1000 \text{ pN}$$

is assumed [150]. The functional form is solved for the force and integrated using

$$U^{\text{hd}}(x^{\text{hdl}}) = \int_0^{x^{\text{hdl}}} dx F(x)$$

to get an approximation of the energy needed to pull the dsDNA to a certain length  $x^{\text{hdl}}$ .

### 5.10.4 Probability Functions

We assume that the energy in the system can be exchanged between the different parts so that the total energy is given by the sum of its compartments

$$U^{\text{tot}} = U^{\text{trp}}(x^{\text{trp}}) + U^{\text{hd}}(x^{\text{hdl}(1)}) + U^{\text{hd}}(x^{\text{hdl}(2)}) + U_c^{\text{rna}}$$

and using a canonical ensemble with fixed average energy, the probability for the observation of a certain configuration is then given by the Boltzman distribution

$$\begin{aligned} p_c(x^{\text{trp}} | \beta, x^{\text{hdl}(1)}, x^{\text{hdl}(2)}) &\propto \exp(-\beta U^{\text{tot}}) \\ &\propto \exp(-\beta U^{\text{trp}}(x^{\text{trp}})) \exp(-\beta U_c^{\text{rna}}) \times \\ &\quad \times \exp(-\beta(U^{\text{hd}}(x^{\text{hdl}(1)}) + U^{\text{hd}}(x^{\text{hdl}(2)}))). \end{aligned}$$

To remove the dependence on the two different handle distances we use that both lengths are related

$$x^{\text{hdl}} \equiv x^{\text{hdl}(1)} + x^{\text{hdl}(2)} = x^{\text{tot}} - x^{\text{trp}} - x_c^{\text{rna}}$$

and integrate over all possible combinations with fixed total handle extension  $x^{\text{hdl}}$ . To simplify the notation, an effective handle energy  $U^{\text{eff}}$  is introduced, defined as

$$\exp(-\beta U^{\text{eff}}(x^{\text{hdl}})) \equiv \int_0^{x^{\text{hdl}}} dx \exp(-\beta(U^{\text{hd}}(x) + U^{\text{hd}}(x^{\text{hdl}} - x)))$$

leading to

$$\begin{aligned} p_c(x^{\text{trp}} | \beta, x^{\text{tot}}) &\propto \exp(-\beta U^{\text{trp}}(x^{\text{trp}})) \exp(-\beta U_c^{\text{rna}}) \times \\ &\quad \times \exp(-\beta(U^{\text{eff}}(x^{\text{tot}} - x^{\text{trp}} - x_c^{\text{rna}}))). \end{aligned}$$

For the final total probability to find the bead in the trap at a certain position  $x^{\text{trp}}$  w.r.t. a given total trap distance of  $x^{\text{tot}}$ , summing over all possible configurations  $c$  yields

$$\begin{aligned} p(x^{\text{trp}} | x^{\text{tot}}) &\propto \sum_c \exp(-\beta U^{\text{trp}}(x^{\text{trp}})) \exp(-\beta U_c^{\text{rna}}) \times \\ &\quad \times \exp(-\beta(U^{\text{eff}}(x^{\text{tot}} - x^{\text{trp}} - x_c^{\text{rna}}))). \end{aligned}$$

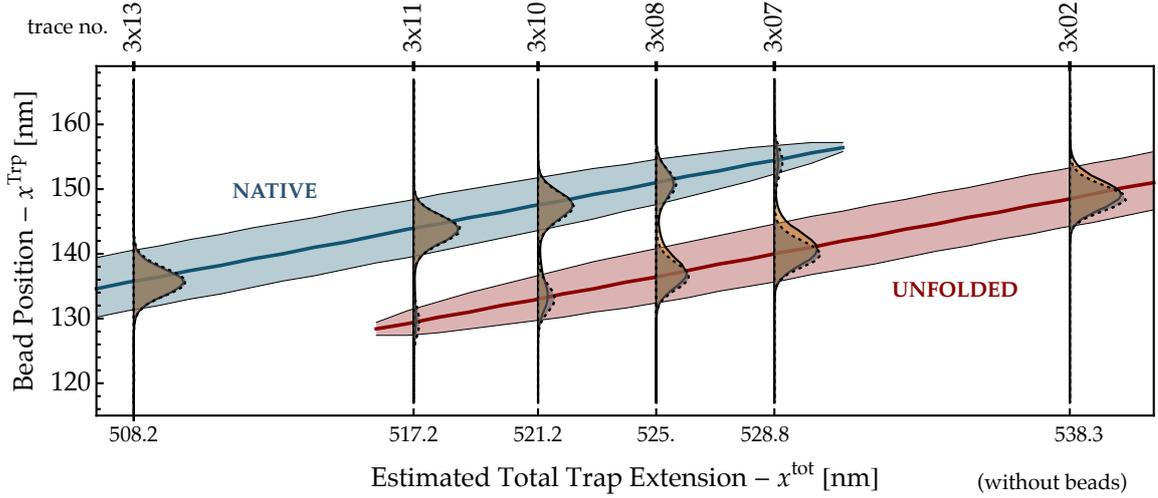
The last expression can be simplified further if a fixed temperature is assumed: Using that the end-to-end distance given in the number of unpaired bases  $n_c$  for conformation index  $c$  can take only finitely many different values  $n_c \in n^{\text{RNA}} \subset \{0, \dots, 56\}$  we can define an energy function

$$\exp(-\beta U_n^{\text{dst}}(n)) \equiv \sum_c \exp(-\beta U_c^{\text{rna}}) \delta(n - n_c)$$

that is end-to-end distance  $n$  dependent. This allows to simplify the solution to

$$\begin{aligned} p(x^{\text{trp}} | x^{\text{tot}}) &\propto \sum_{n \in n^{\text{RNA}}} \exp(-\beta U_n^{\text{dst}}(n)) \exp(-\beta U^{\text{trp}}(x^{\text{trp}})) \times \\ &\quad \times \exp(-\beta(U^{\text{eff}}(x^{\text{tot}} - x^{\text{trp}} - x^{\text{rna}}(n)))), \quad (5.10.2) \end{aligned}$$

which is finally a sum over all possible end-to-end distances  $n^{\text{RNA}}$  and not over the exponentially growing number of configurations. For the



**Figure 5.8 – Equilibrium Predictions of the RNA Model for different total trap positions** (dashed lines/gray). The Model shows the expected qualitative behavior for the relative bead displacement: Increasing the total trap-distance (trap position) also increases the strain on the RNA hairpin. At  $x^{\text{tot}} > 515$  nm the unfolded states (red) appear and until  $x^{\text{tot}} < 530$  nm also the native conformations (blue) are present. Shown are only positions that match a certain experimental trajectory (identification on top). Histograms computed from the experimental traces (solid lines/yellow) used to match the total trap distance  $x^{\text{tot}}$ . The numbers/ids on top refer to the original numbering in the tweezer experiment (e.g. 3x08 is trace no. 3 and trap setting no. 8 - see appendix D.2 for details).

end-to-end distance function  $x^{\text{rna}}(n)$  we used proportionality factor of 0.34 nm/base. By precomputing  $U_n^{\text{dst}}(\beta)$ , the probability  $p(x^{\text{trp}} | x^{\text{tot}})$  can be computed quickly even for very large sets of configurations. For practical purposes, the sum in  $U_n^{\text{dst}}(\beta)$  is again approximated by all configurations with

$$U_c^{\text{rna}} \delta(n - n_c) \geq \min_c U_c^{\text{rna}} \delta(n - n_c) - \Delta U$$

which effectively removes all configurations being a factor of  $e^{-\Delta U}$  less probable than the most likely configuration for a given end-to-end distance  $n$ . Setting  $\Delta U = 10 \text{ kcal mol}^{-1}$  leaves a remaining 17,272 configurations.

### 5.10.5 Predicted bead position histograms

Using the RNA model we predicted stationary probabilities  $p(x^{\text{trp}} | x^{\text{tot}})$  to observe a certain trap dislocation  $x^{\text{trp}}$  given a specific total trap extension  $x^{\text{tot}}$  at room temperature 293 K. Figure 5.8 shows the analytical results as gray shaded regions with dashed lines and the matching histograms from the 6 experimental traces in yellow with solid lines. The trace-specific total trap extensions  $x^{\text{trp}}$  have been fitted manually to match the experimental prediction best and in a

Trace ID	3x02	3x07	3x08	3x10	3x11	3x13
$x^{\text{tot}}$ [nm] w/o beads	508.2	521.2	524.0	528.8	531.3	538.3
$x^{\text{tot}}$ [nm] with beads	3,158.2	3,171.2	3,174.0	3,178.8	3,181.3	3,188.3

**Table 5.1 – Estimated total trap extensions** Estimated total trap extensions  $x^{\text{tot}}$  for each of the 10 recorded traces. The distance is given without the additional two bead radii of 2650 nm in total as it will be done in the calculations. The trace ids match the experimental setup and refer to traces recorded from fiber no. 3 (see appendix (D.2)).

range between 505 nm and 540 nm either folded and/or unfolded structures are present. See Table 5.1 and Figure 5.8 for exact numbers.

With increasing total distance  $x^{\text{tot}}$  both, the folded and unfolded structures, move towards higher bead dislocations  $x^{\text{trp}}$ , which is expected and is due to the weak trap constant  $k^{\text{Trap}}$ . Increasing the pulling force will mainly move the bead in the trap in the same direction until the opposing force is large enough to cause the hairpin to unfold. Therefore the unfolded and thus more extended state corresponds to a lower bead displacement from the trap center  $x^{\text{trp}}$ . To improve the agreement between experimental histograms and the model prediction we manually reduced the trap constant from an initially experiment-derived guessed value of

$$k^{\text{Trap}} = 0.100 \text{ pN/nm}$$

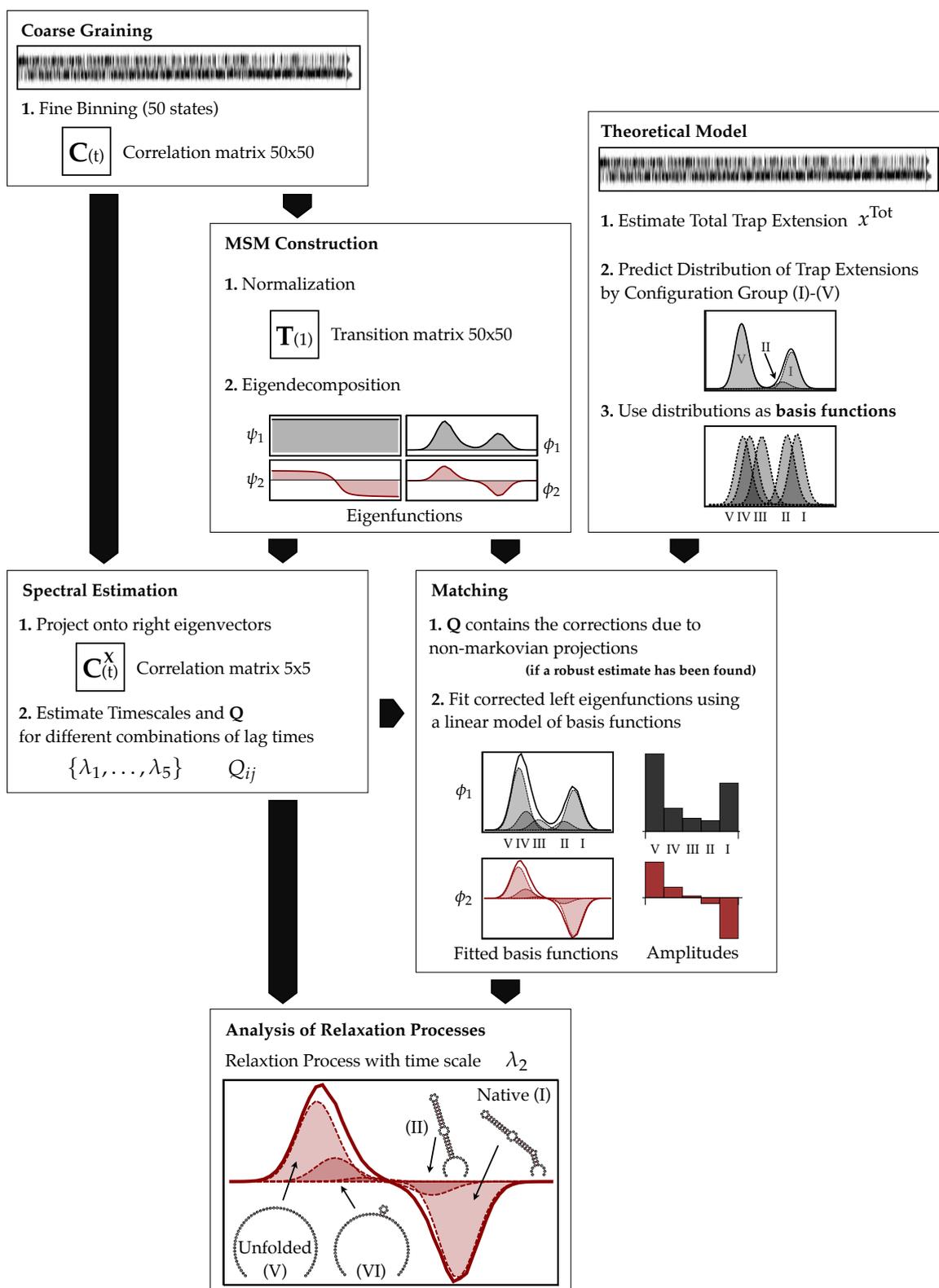
by 4% to the finally used value of

$$k^{\text{Trap}} = 0.096 \text{ pN/nm}$$

until the shape of both histograms matched best. All other model constants were kept as the typical values given in the previous sections. Considering that only a single parameter had to be adjusted the agreement between model and experiment is very good. We expect that here is still room for improvement to adjust all parameters to get even better results.

#### 5.11 RECONCILING FORCE PROBE EXPERIMENT AND RNA MODEL

So far we only compared the stationary probability distribution of bead dislocations  $p(x^{\text{trp}} | x^{\text{tot}})$  and we want to proceed with the following dynamical analysis as illustrated in Figure 5.9: Starting from the previously estimated total trap extension  $x^{\text{tot}}$  for a specific trace the introduced model can predict the stationary bead dislocation separately for RNA configurations that exhibit a certain end-to-end distance  $x^{\text{trp}}$  using Eq. 5.10.2. Without any simplification the 56-bases of the RNA hairpin can have 57 different end-to-end distances (0 to 56) which we grouped into 5 clusters to reduce the complexity of the model. To define the clustering we examined all 57 end-to-end distances in Eq. 5.10.2 and computed their relative contribution to the



**Figure 5.9 – Schematic analysis of relaxation processes** using experimental estimations (middle), theoretical predictions (right) and the relation to spectral estimation (left). On the experimental side a MSM is generated from the trajectory and dominant eigenvectors are computed that are used for spectral estimation and as indicators for the dynamics in the system. On the theoretical side the total extension  $x^{\text{tot}}$  is estimated that allows then to predict the probability distribution for this extension. From this the 5 dominant groups (I, ..., V) of end-to-end distances are computed which form the basis for the matching onto the dynamics encoded in the eigenvectors from the experimental part. Together with spectral estimation these three parts together allow an estimation of the time scales for each relaxation process and the associated structural changes between the configurational groups (I-V). See Figures 5.10, (5.11) and 5.6 for final estimation results.

	$x^{\text{rna}}$ [bases]	max contrib.	$x^{\text{rna}}$ [bases]	max contrib.
a)	56	98,99%	40	0,47%
	9	90,94%	20	0,17%
	18	6,43%	38	0,15%
	11	3,66%	50	0,10%
	16	2,68%	rest	<0,10%

b)	group	I	II	III	IV	V
	$x^{\text{rna}}$ [bases]	{9,11}	{16,18,20}	{38,40}	{50}	{56}

**Table 5.2 – Definition of the 5 relevant end-to-end distance groups** used in the theoretical model for the dynamical analysis. a) **Maximal relative contribution to the total distribution of predicted bead displacements**  $x^{\text{trp}}$  from RNA configurations with a distinct end-to-end distance given in the number of free bases (see Figure 5.7 for an example). b) **Definition of the 5 relevant end-to-end distance groups** based upon the results in Tab. a). All configurations with a maximal relative contribution of more than 0.1% were used and distances of 2 or less bases were considered indistinguishable and hence put in the same cluster.

stationary distribution over a range of  $x^{\text{tot}} \in [505 \text{ nm}, 540 \text{ nm}]$ . We then selected all end-to-end distances that contributed at some  $x^{\text{tot}}$  at least 0.1% to the total displacement distribution  $p(x^{\text{trp}} | x^{\text{tot}})$  when a binning into 50 states was used. This left 9 relevant end-to-end distances out of the 57 possible ones.

The 5 groups were finally chosen under the assumption that differences of 2 bases or less are indistinguishable due to the resolution of the experiment, which resulted in 5 groups that are indicated by roman numerals and are ordered with increasing end-to-end distance from (I) folded to (V) unfolded. See Tab. 5.2 for concrete numbers and Figure 5.7 for a list for the dominant configurations for each of the 5 groups. These five subgroups cover at least 99.69% of the total probability in the bead displacement distributions, which we consider to be a good approximation of the full 57 end-to-end distances model.

The next step is to compute a basis  $p_i^{\text{Basis}}(x^{\text{trp}} | x^{\text{tot}})$  for the distribution based on the five subgroups

$$p_i^{\text{Basis}}(x^{\text{trp}} | x^{\text{tot}}) \propto \sum_{n \in \text{group}(i)} \exp(-\beta U_n^{\text{dst}}(\beta)) \exp(-\beta U^{\text{trp}}(x^{\text{trp}})) \times \exp(-\beta(U^{\text{eff}}(x^{\text{tot}} - x^{\text{trp}} - x^{\text{rna}}))),$$

with  $\text{group}(i)$  containing only the distances of subgroup  $i$  (see Tab. 5.2, b). From now on we will assume that the stationary probability distribution of relative bead positions found in the experiment as well as the dynamics of all relaxation processes (i.e. the eigenfunctions) can

be approximated by a linear combination of these five basis functions. Note, that due to the nature of the physical setup the basis functions move and change shape with increasing total trap extension  $x^{\text{tot}}$ . Expressed in a more illustrative way, we explicitly assume now that the dominant dynamics can be expressed as transitions between the five subgroups of end-to-end distances.

To compute the eigenfunctions that express the direction of the relaxation processes we constructed a MSM using the native (sampling) lagtime  $\tau = \tau_0 = 0.020$  ms. These functions have been used before as the projection in the spectral estimation procedure and are now considered an approximation to the correct projected eigenfunctions. In principle we could now use the  $\mathbf{Q}$  matrix to correct for the fact that the estimated eigenfunctions  $\hat{\psi}_i$  are by construction orthogonal, while the actual projected eigenfunctions are not. In this analysis we did not include this additional correction since we had no robust method to combine the results from the different estimations into a single estimate. This is an important point that needs to be investigated further.

What we would finally like to show is, that the experimentally observed dynamics of the dominant relaxation processes – the ones contained in the dominant eigenvectors  $\hat{\psi}_i$  – can be expressed in terms of a basis that is derived from the dominant structures of the theoretical RNA model. In this case we can argue that we can actually assign structural changes to experimental relaxation process provided that our model is a good approximation to the real dynamics (see Figure 5.9 bottom).

For the actual comparison, the basis functions  $p_i^{\text{Basis}}$  were fitted to the 4 dominant (left) eigenvectors  $\hat{\psi}_i$  using a least-squares fit using an algorithm from Mathematica (command FindMinimum). As seen in Figure 5.10, the agreement between the experimental distributions/eigenvectors (solid) and the linear combination using the basis functions  $p_i^{\text{Basis}}$  (dashed) are visually quite good, especially for the stationary distribution and the two slowest processes. We should note, that in principle using a sufficiently large set of basis functions any function can be fitted arbitrarily well. In our case we wanted to approximate 4 eigenfunctions that have been discretized at 50 spatial points by a set of 5 basis functions that are solely computed from a theoretical model. No further optimization of these basis functions apart from a linear recombination was made and hence the argument of over-fitting does not hold here. The agreement for process number 4 (green) is less accurate and seems to involve additional states. Indeed, the matching can be improved by adding the next relevant configurations w.r.t. to end-to-end distance following the above scheme.

Figure 5.11 shows the amplitudes of the fitted basis functions and hence the actual linear combination of basis functions used in the dynamical analysis. The aim is to see if the structure of the eigen-

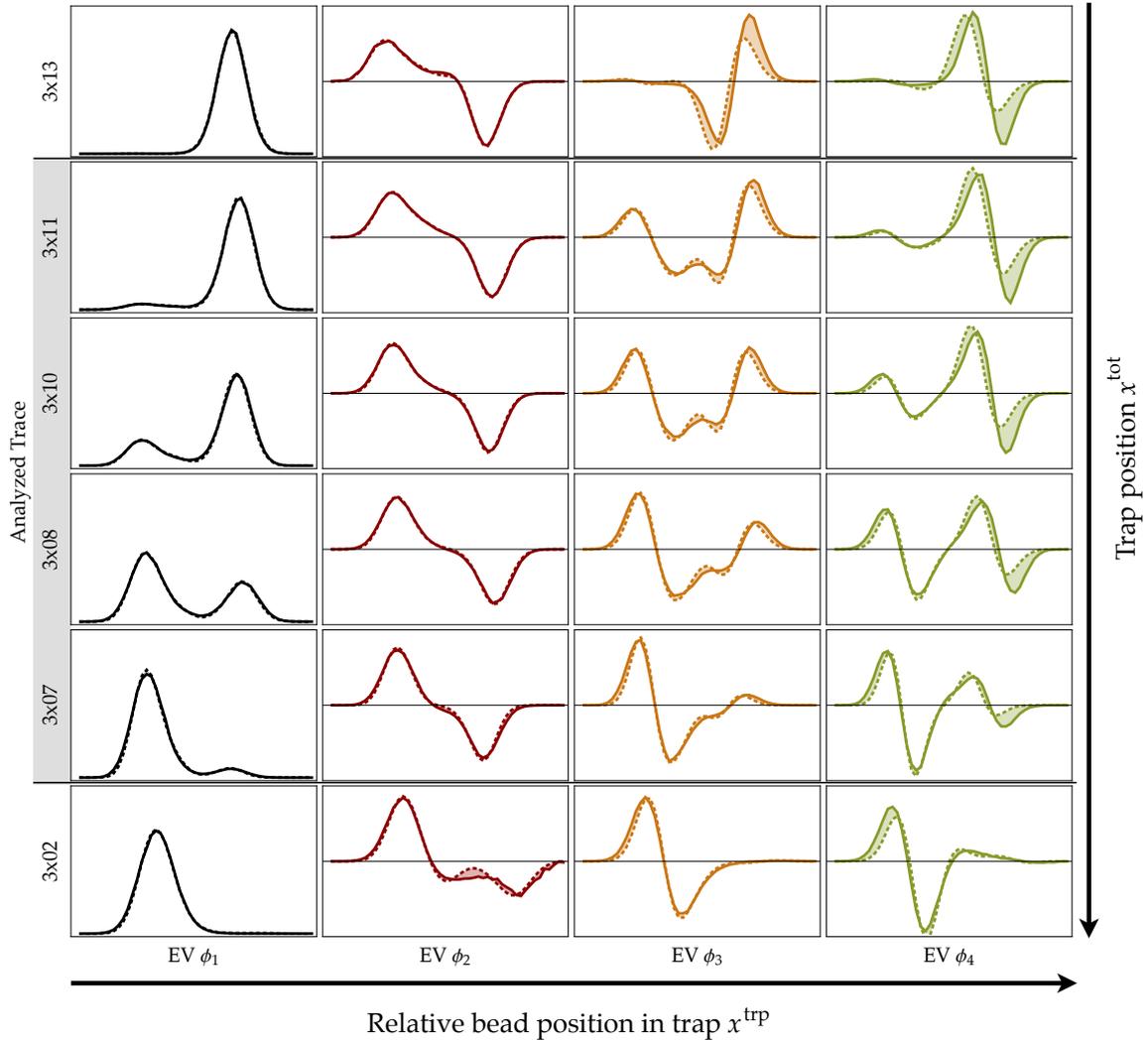
vectors in the theoretical basis is similar among different traces and if we can assign a structural change to the different relaxation processes. All processes show the same structure (positive/negative regions) for each trace suggesting that the assignment of a relaxation process to a time scale is the same between the different traces. The four traces in the middle (marked with a gray bar) show a clear 2-state histogram and we expect the dominant processes to be more distinct there compared to the traces where only a single state is observed (trace 3x02 and 3x13). The plot further illustrates that the slowest process (red) can be interpreted as a transition between the native (Group I) and the unfolded configurations (Group V). The 3rd eigenvector (yellow) involves transitions between the intermediate states (Group II-IV) and the fastest dominant process, Eigenvector 4 (green), finally represents transitions within each macro state (Groups I+II and Groups IV+V) of the visually apparent (2-state) nature of the hairpin dynamics. This ordering of time scales of processes is what would be expected for a simple folding mechanism: The fast processes ( $\lambda_4$ ) are within the macro (visible) states while the transition between macrostates is slower ( $\lambda_2$ ).

## 5.12 SUMMARY AND DISCUSSION

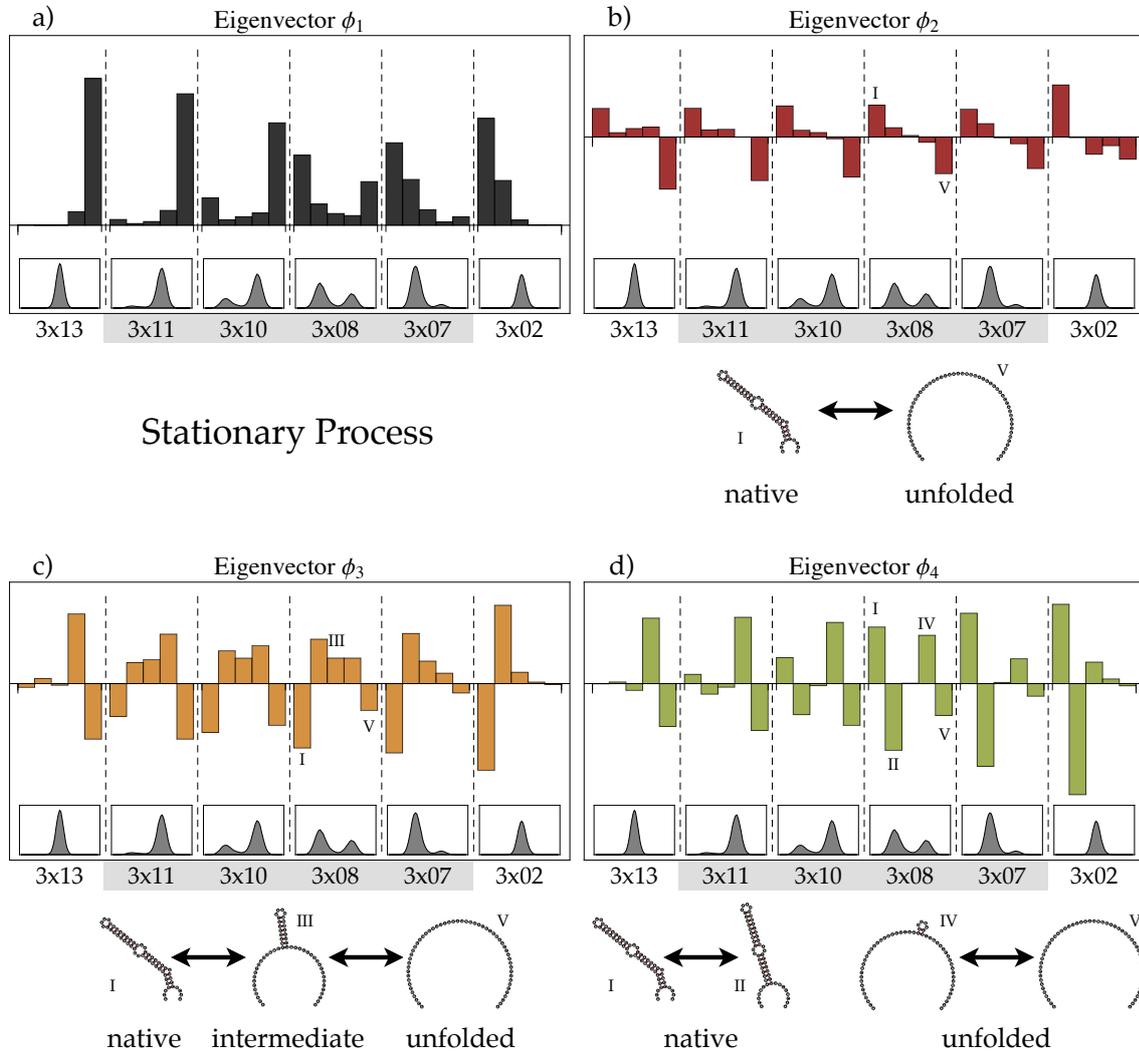
In the first part of this chapter we have presented a new approach to estimate time scales and the corresponding projected eigenvectors of an underlying Markov Model from very low-dimensional projected time series. The method requires information from observations collected at at least two different lag times utilizing ideas from a generalized eigensystem decomposition and can partially circumvent the projection error present in MSM time scale estimations. This places the method in the group of multi- $\tau$  estimators with improved convergence properties.

A basic sensitivity and error analysis was conducted that showed a slightly higher sensitivity to the initial correlations compared to the MSM estimation due to the dependence on more initial values which should be reduced by optimization strategies incorporating observations from more than the minimal required number of two different lag times. The question of robustness against statistical fluctuations is still a major point in the ongoing investigation. We assume that a scheme, starting with an estimation of only the slowest process and iteratively increasing the number of processes while keeping the already estimated ones fixed, might improve stability. This would also allow to optimally incorporate the different lag time regions where a process is observable.

From a theoretical point of view, another possible solution would be to reformulate the spectral estimation in terms of a variational principle with a suitable target function that takes a minimum once



**Figure 5.10 – Approximation of eigenvectors estimations for various experimental traces** indicated by the trace id on the left. The gray bar indicates traces that show a clear 2-state histogram and are therefore expected to show more dynamical interaction. Solid lines refer to the MSM based eigenvector, dashed lines to the approximation by the theoretical basis. Colors indicate the different relaxation processes: stationary  $\hat{\phi}_1$  (black), slowest relaxation  $\hat{\phi}_2$  (red),  $\hat{\phi}_3$  (yellow),  $\hat{\phi}_4$  (green). The agreement is quite good especially for the three slowest processes and the traces that show 2-state in their histogram.



**Figure 5.11 – Eigenvector approximation by theoretical basis** Shown are the amplitudes used to fit the four dominant eigenvectors  $\{\hat{\psi}_1, \dots, \hat{\psi}_4\}$  using the basis functions  $\{p_1^{\text{Basis}}, \dots, p_i^{\text{Basis}}\}$ . Each of the four subplots presents the presence of the basis functions for the six analyzed experimental traces. Below each plot a schematic interpretation of the relaxation process is depicted. (a) The stationary process, (b) process 2 (red) moves between the native (I) and the unfolded configurations (V), (c) process 3 (yellow) involves intermediate states (II-IV), and (d) process 4 (green) indicates transitions within the two visual macrostates (I+II and IV+V).

an optimal solution/projection has been found. This would be similar to the Ritz method that can give the optimal projection provided that the projections are chosen to be orthogonal. In the ideal case we would like to find a target function that is independent of the orthogonality and uses in addition correlations matrices at multiple points in time to reduce perturbations from statistical uncertainties.

In a second step, the method was applied to the artificial 4-state system that was introduced in chapter 2. The results show much improved convergence to the correct time scales compared to ITS from MSM. Especially, once the fast processes can be neglected, the error practically vanishes. This was done for a variety of different projections which differ significantly only in their convergence behavior in the lag time region where the fast processes are still present. This is in accordance with the theoretical findings that the method is exact in the slow time scale region while it still depends on the choice of projections if the lag time  $\tau$  is chosen too small. For projections that eliminate the the fast processes the estimation is exact over the full range of lag times as it is used e.g. in PCCA. Thus, as long as statistical uncertainties are not an issue this method provides much faster convergence to the true time scales compared to the construction of a MSM and compute Implied Time Scale from it. Hence, we expect that spectral estimation can be a replacement to the implied time scales estimation that is widely used as being explained in chapter 2.4.3.

In the third part, the effectiveness of the new method was demonstrated on experimental time series from an optical tweezer experiment of a 56-base ssRNA hairpin. We successfully estimated time scales for the four dominant processes that showed a consistent change of the estimated time scales over a set of six independently collected time series ranging from the totally closed to the fully extended state. So far the estimations were done for a variety of different sets of lag times using the presented optimization scheme which allowed a qualitative analysis of the variance that is caused by the statistical influences. Under the influence of statistical uncertainty the optimization suffered from convergence problems which needs to be addressed in further investigations. The previous idea of an iterative and lag time dependent estimation might increase the chances of convergence but also a more robust optimization implementation is necessary. Better initial estimation parameters might also improve the estimation and reduce the necessary number of iterations.

Since we base the estimation on correlation functions, it is reasonable to use methods that assure a good convergence for the correlations as well. One way to improve this is to use dynamical reweighting [4, 3] as we demonstrated in chapter 4. This might actually prove to be a powerful combination since dynamical reweighting can reduce the statistical uncertainty which is so far the crucial point in the spectral estimation.

In the last part, we related the experimental findings to conformational changes in the structure of the RNA hairpin. For this a computational model for a RNA hairpin in an optical trap was build upon the RNA structure prediction from the Vienna RNA model web service. The model was based solely on physical considerations and allowed very good predictions of the stationary distributions observed in the experimental time series. The necessary parameters were taken to be widely accepted averages and only the trap constant for the harmonic laser trap had to be adjusted by a reduction of 4%, which seemed necessary for the model to reproduce the observation. The matching is excellent for the closed states and differ slightly in the transition region and the open states. This might be due to an increase of non-harmonicity in the force of the trap with increasing total trap distance. Also, we expect that the RNA structure prediction was optimized to predict the low-energy states correctly and thus the variation in the higher energy states could be a trade-off of the simplicity of the model. We also have to take into account that the trajectory might not be converged although in this case we would expect to miss certain conformations instead of potentially high energy states being overpopulated. Nevertheless the agreement between the theoretical model and the experimental observations is better than we had expected.

The RNA model was used to connect the structural information present in the theoretical model to the dynamics in the experimental time series. For this we estimated five relevant groups of configurations that were clustered by their end-to-end distance. The contribution of these groups to the distribution of bead displacements was used to find a set of five basis functions. This basis was assumed to be the building blocks of the dominant dynamics, i.e. the slow relaxations processes move between these five subgroups of configurations. It was then mapped onto the eigenfunctions from a MSM model that was estimated at the native sampling lag time which finally allowed for an intuitive interpretation of the relaxation processes.

The agreements were very good for the four dominant (including the stationary) processes with minor deviations in the folded part of the fastest of these processes. The particular combination of relaxation processes suggests a slow transition between the two apparent macro states that a visible in the histogram with few intermediates (similar to a 2-state folder), while the faster processes describe transitions within the groups of folded or unfolded states. We assume that our choice of five basis groups is not fine enough and thus responsible for the insufficient good approximation in the fast processes. A more thorough investigation and especially an application to traces with lower noise should provide more insight into this matter. In conclusion, the presented connection between the experimental data

and the RNA model can give additional insight into the observed processes and their internal dynamics.



---

## SUMMARY & OUTLOOK

---

### STATE-OF-THE-ART

Recapitulating the content of the introduction and theory chapter 2 we realize that the theory of Markov models – in their ability to provide means to describe the relevant dynamics of biomolecules – has matured into a well-developed and widely used theory. There exists the broad mathematical foundation from Markov processes, rigorous error bounds, and construction strategies that can be proven to be optimal under certain conditions. Aside from the theory, there is a variety of successful applications ranging from the construction of Markov State Model (MSM) for simple and well-studied test cases [65] up to the construction of models that allow to span time scales much larger than the actual simulation time possible for peptides of that size [35]. Thus, it is possible to robustly generate MSMs from simulations, validate their predictions and compute a variety of important key properties from it.

Acknowledging this development, we have not yet paid the necessary attention to the question of how to obtain the information needed to parametrize our models to the required accuracy. It turns out that for systems of relevant sizes the sampling problem poses the main issue, i.e. the exponential growth of state space with increasing system size and the existence of high barriers slow down the exploration in the state space. Hence, for large systems, the necessary simulation time exceeds what is computationally affordable or even possible. Here, an important question comes up, aside from the obvious ways to improve the existing methods: Can MSMs provide a way to speed up the discovery of the relevant phase space parts compared to what is possible from a single long simulation? This problem was one of the original issues that were to be addressed by the application of MSMs. Concretely, we would like to know whether the locality of the conditional jump probabilities in the transition matrix can be used to formulate a strategy that will reduce the simulation time for molecular dynamics simulation by providing means to patch information from several short trajectories into an improved model. It was already successfully demonstrated that this is possible to artificially increase

the affordable simulation time to estimate the equilibrium ensemble for a medium-sized peptide [35].

#### ENHANCED SAMPLING USING SHORT TRAJECTORIES

The prospect that Markov models could help to circumvent or mitigate the sampling problem by utilizing a large set of short trajectories starting from different conformations conventionally only rarely visited in equilibrium, raises the question of how the relevant starting conformations can be found. This question is not specific to Markov model analyses, and it is likely that in this stage biased sampling methods such as meta-dynamics [152], conformational flooding [153], umbrella sampling [154], targeted Molecular Dynamics (MD) [155], replica-exchange MD [113], or pathway methods [46] will be useful to generate an initial exploration of conformation space from which short equilibrium simulations can then be launched. Once the relevant conformations have been found and a good discretization has been obtained, it is clear that the uncertainty estimates of the Markov model can be exploited in order to pick starting points of subsequent simulations so as to adaptively reduce the uncertainty in the parametrization of the Markov model. Using data of a parallel tempering simulation this was already demonstrated [114] and this work has extended this idea in chapter 4 to the combined usage of simulations from multiple ensembles. Still, these models require manual intervention such as the decision for a suitable discretization or the number of dominant processes. The next step to an automated exploring of state space has yet to be achieved and the methods and findings presented in chapter 2 and 4 can contribute to the necessary basis for this.

#### STEPS TOWARDS ADAPTIVE SAMPLING STRATEGIES

We realize that the fast and robust parametrization of the Markov model is the missing key problem, and not the optimization and computational steps once a Markov model has been generated. Assuming that both of these parts function reliably, then – at least theoretically – the available computational resources could be chosen to maximize the convergence rate of some user specified observable property instead of the goal to explore the full state space as fast as possible. One could for example choose to estimate a specific eigenvalue-eigenvector pair to a certain accuracy as fast as possible while neglecting all parts of the configurational space that are not important in this case.

For such a scheme to work, reliable error estimations for these properties are indispensable. One often used property is the committor since it provides an optimal reaction coordinate and is a basis for

Transition Path Theory (TPT). In chapter 3 we have derived algebraic expressions for the necessary variation in the committor. A second even more used key property is the set of dominant relaxation time scales of the system, the estimation of which is one of the major objectives. As we have demonstrated in chapter 5 there exist ways to circumvent the projection error for some of these properties that allows for an increased accuracy of estimation. This has not yet been used in the context of Markov models but it might represent a replacement for the much used Implied Time Scale (ITS) of MSMs and can improve the convergence speed of future adaptive sampling schemes. This will of course take time, since not only the mathematical foundations need to be developed but also an integration into existing simulation frameworks need to be achieved. Still, the basis exists and the necessary steps towards enhanced sampling are already being developed [127, 60, 61].

#### BRIDGING THE GAP BETWEEN SIMULATION AND EXPERIMENT

The origin of most simulations is that the implications drawn from it are in some sense a mirror of reality and can provide a more detailed insight into the dynamics of systems that are not or only partially accessible by other, experimental or theoretical, means. Hence, simulations can be seen as an in-between of theory and experiment. While the sampling problem still bounds the accessible time scales in simulation from above, there exists a converse lower bound to the times that can be investigated experimentally. An overlap and hence a direct comparison of simulation and experiment is thus rarely possible, especially in biologically relevant cases. Since we base our Markov state models on simulations rather than theoretical considerations, it is desirable to derive a connection between the properties of Markov models and the observations from experiments. Here, the results from chapter 5 can help to improve estimations and thus serve as a bridge between experimental estimations and the thoroughly investigated Markov models from simulations and therefore eventually contribute to their direct comparison.

#### APPLICATIONS FOR OTHER KINETIC MODELS

The type of Markov models investigated here, i.e., transition matrix based kinetics models between discrete state partitions of configuration space, must be viewed as one aspect within a family of conformation dynamics approaches. Rate matrix or Master equation models [44, 73, 74] are very close in spirit, and we have mentioned connections to these models (see section 2.3), making most of our present results available to these models as well. Recently, an alternative approach [44] has been proposed to obtain coarse-grained Markov or

Master equation models based on a non-complete partition of state space that avoids to finely discretize the transition region. It is shown in [78] that the presented analyses of the discretization error can be applied to this approach as well, only that here the eigenfunctions on the non-resolved parts of state space are effectively replaced by an interpolation based on committor functions between core sets[108]. Generating Markov or Master equation models based on rate models from an exploration of the stationary points of the energy landscape is an approach that has great tradition [40] and has been particularly successful in the analysis of Lennard-Jones or water clusters [156, 40]. These models are not concerned with the same estimation problems as the present Markov models, as they are built from rate-theory based estimates (such as transition state theory) of state-to-state transition rates between the stationary points of the energy landscape, and not from trajectory statistics. However, they necessarily share the same concerns of making a discretization error by aggregating points of continuous state space into discrete model states. In a wider sense, approaches that use MD to parametrize effective stochastic equations, such as Langevin dynamics [91, 157, 158], also induce models of the ensemble dynamics, such as Fokker-Planck type models. These ensemble dynamics models generally share the advantages of Markov models over traditional MD analyses that have been discussed in the introduction. The specific advantage of Markov models is that they are on one hand asymptotically exact both in terms of discretization and estimator quality (see Sec. 2.4 and 2.6), and on the other hand very simply compared to models that in some way include memory. In addition, the concept of projected Markov models from chapter 5 provides an alternative view on the virtual memory introduced by projections.

As MSMs allow the whole arsenal of Markov chain theory to be readily accessed, the functional relationship between Markov models and most interesting molecular properties or observables has been worked out already [60, 58, 35, 43, 55, 63], and often has a simple and straightforwardly interpretable form. Given these advantages we expect that the popularity of Markov or similar models for the modeling of molecular kinetics will keep increasing. As we have demonstrated, the simulation and estimation capabilities of Markov State Model have developed greatly until now and will also make a relevant contribution to the understanding of molecular kinetics in the future.

# A

---

## NOTATION AND SYMBOLS

---

To simplify the readability, all symbols commonly used throughout the thesis are listed in Tab. A.1 and Tab. A.2. Any deviations from this are mentioned and repeated where necessary or helpful. Ornaments (hats, dagger, etc.) associated with symbols mostly refer to specific attributes listed in Tab. A.3 and a summary about notational style is given in Tab. A.4. To shorten the notation (especially for linear algebra) and ensure a better readability a few abbreviations and conventions have proven helpful:

1. Indices of vectors and matrices start with one.
2. Indices for vectors, matrices and higher order object are indexed with subscript  $T_{ij}$ .
3. If the notation is long or ambiguous, []-brackets are used to indicate the use of indices, e.g.  $T_{[i,x(t)]}$ .
4. If indices are used superscript they are always in brackets to distinguish them from the power operation, e.g.  $x^k$  vs.  $x^{[k]}$ .
5. A set can be used as an index, indicating, that the object is reduced to a sub-object containing only the given indices, e.g.  $\mathbf{T}_{\{1,2,3\},\{1,2,3\}}$  would be the sub-matrix taking the first three rows and columns,  $\mathbf{q}_{[A]}$  indicate the vector  $\mathbf{q}$  restricted to the entries listed in set A.
6. Sets are indicated in curly {}-brackets, e.g.  $A = \{1, 3, 4, 8\}$ .
7. A colon : (as used in Matlab) can be used to shorten ranges of indices, e.g.  $\mathbf{T}_{\{1,2,3\},\{1,2,3\}} = \mathbf{T}_{[1:3,1:3]}$ .
8. A dot  $\cdot$  refers to all entries in an index, this way a row or a column can be selected, e.g.  $\mathbf{T}_{[i,\cdot]}$  refers to the  $i$ -th row of the matrix, while  $\mathbf{T}_{[\cdot,j]}$  is the  $j$ -th column. Both objects are considered a (column) vector now and are treated as such.
9. The indices are always taken last, after inversion, the transpose, etc. , e.g.  $\mathbf{T}_{[i,\cdot]}^T$  is the  $i$ -th row of the transposed of  $\mathbf{T}$  as a vector
10. If an equation is split into multiple lines, a multiplication is indicated using  $\times$

Symbol	Element	Meaning
$\Omega$	$x \in \Omega$	state space
$\tau$		lag time, time resolution of the model.
$\mathbf{x}$	$x_t, x_k \in \Omega$	time series in $\Omega$ continuous in $t$ , discrete in $k$
$\mathbf{p}_t$	$\mathbf{p}_t(x) \in \mathbb{R}^+$	continuous (in state space) probability density.
$\mathbf{p}_\tau$	$\mathbf{p}_\tau(x, y) \in \mathbb{R}_+$	transition probability density
$\mathbf{Q}(\tau)$		propagator for lagtime $\tau$
$\mathbf{T}(\tau)$		transfer operator for lagtime $\tau$
$\Pi$		reweighting operator for dual elements in $L^2_\pi$
$\lambda_i$	$\lambda_i \in [-1, 1]$	eigenvalues
$\pi$	$\pi(x) \in \mathbb{R}^+$	stationary density
$\psi_i$	$\psi_i(x) \in \mathbb{R}$	eigenfunctions of $\mathcal{T}(\tau)$
$\phi_i$	$\phi_i(x) \in \mathbb{R}$	density-weighted eigenfunctions of $\mathcal{T}(\tau)$
$\delta(\cdot)$	$\delta(x \in A) \in \{0, 1\}$	indicator function, equals 1 if $x \in A$ and 0 else
$\chi_i$	$\chi_i(x) \in [0, 1]$	degree of membership of $x$ to discrete state $i$
$\xi_i$	$\xi_i(x) \in \mathbb{R}$	value of state $x$ of observable/projection $i$

**Table A.1 – Commonly used symbols** throughout the thesis

Symbol	Elements	Name
$\mathbf{T}(\tau)$	$T_{ij}(\tau) \in [0, 1]$	transition matrix of conditional jump probabilities
$\mathbf{C}(\tau)$	$C_{ij}(\tau) \in \mathbb{R}$	correlation matrix/matrix of absolute jump probabilities
$\mathbf{p}(\tau)$	$p_i(\tau) \in [0, 1]$	probability distribution
$\mathbf{\Pi}$	$\Pi_{ii} \in [0, 1]$	diagonal matrix of probabilities $\Pi_{ii} \equiv \pi_i$
$\mathbf{R}$	$\mathbf{r}_i$	matrix of right eigenvectors $\mathbf{r}_i$ in columns of $\mathbf{T}$
$\mathbf{L}$	$\mathbf{l}_i$	matrix of left eigenvectors $\mathbf{l}_i$ in columns of $\mathbf{T}$
$\mathbf{\Lambda}(\tau)$	$\Lambda_{ii} \in [-1, 1]$	diagonal matrix of eigenvalues $\lambda_i$ of $\mathbf{T}$
$\mathbf{X}$	$X_{ij} \in [0, 1]$	matrix of projection/membership functions $\chi_i(x)$ in columns
$\mathbf{C}^X$	$C_{ij}^X \in \mathbb{R}$	observation Correlation Matrix
$\mathbf{Q}$	$Q_{ij} \in \mathbb{R}$	projected left eigenvector matrix
$\mathbf{Z}(\tau)$	$Z_{ij} \in \mathbb{R}_+$	transition count matrix (row-dominant)

**Table A.2 – Transition Matrix Notation** Symbols used for discrete time, discrete state space Markov State Model ([MSM](#))

Ornament	Name	Meaning
$\hat{x}$	hat	estimated values
$\bar{x}$	bar	mean value, e.g. ensemble average
$\mathbf{A}^T$	upper T	transpose of a matrix
$\mathbf{A}^{-1}$	upper -1	inverse of a matrix
$\mathbf{A}^{-T}$	upper -T	inverse transposed of a matrix
$A^\dagger$	upper dagger	adjoint operator (w.r.t. the invariant measure)

**Table A.3** – Symbol ornaments as used in the thesis.

Notation	Name	Style
$x, y, z$	scalar (continuous)	lower-case, italics
$\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{z}$	vector	lower-case, bold
$\mathbf{X}, \mathbf{C}, \mathbf{T}$	matrices	upper-case, bold
$i, j, k, l, n, m$	integer indices/variables	
$N, M, T, \beta$	global parameters	
$\Omega, \mathbb{N}, \mathbb{R}, A, B$	sets	script
$A, Q, P$	operators	upper-case, non-italics
$\langle f, g \rangle := \int \mathbf{a}_x \pi(x) f(x) g(y)$	scalar product (continuous)	
$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_i \pi_i a_i b_i$	scalar product (discrete)	

**Table A.4** – Notation Symbols used for general types of mathematical objects



# B

---

## MULTI ENSEMBLE ESTIMATION

---

### B.1 EFFICIENT SOLUTION OF THE SELF-CONSISTENT EQUATIONS FOR CANONICAL DISTRIBUTION OF HAMILTONIAN TRAJECTORIES

For the case of a canonical distribution of Hamiltonian trajectories, the normalization constants  $\hat{Z}_k$  or alternatively the dimensionless free energies  $\hat{f}_i \equiv -\ln \hat{Z}_i$  are defined through a set of  $K$  coupled nonlinear equations

$$\hat{f}_i = -\ln \sum_{n=1}^N \left( \sum_{k=1}^K N_k \exp[\hat{f}_k - (\beta_k - \beta_i)E^{[n]}] \right)^{-1} \quad (\text{B.1.1})$$

where all symbols as defined as for Eq. (4.2.3). Any numerically stable method for solving a set of coupled nonlinear equations can, in principle, be used to obtain the  $\hat{f}_i$ . A scheme for solving a more general form of these equations by self-consistent iteration or Newton-Raphson is described in Appendix C of [120].

Because of the structure of this specific case, we can rapidly obtain a close initial guess for the  $\hat{f}_i$  by using a form inspired by the weighted histogram analysis method (WHAM) [118]. By instead constructing  $M$  bins in the total energy  $E$  spanning a range  $(E^{\min}, E^{\max})$ , we can approximate Eq. (B.1.1) with a sum over histograms (as in Eqs. 19–20 of [118]):

$$\hat{f}_i^{[j+1]} = -\ln \sum_{m=1}^M H_m \left[ \sum_{k=1}^K N_k \exp[\hat{f}_k^{[j]} - (\beta_k - \beta_i)E_m] \right]^{-1}$$

where  $H_m$  denotes the number of samples  $E_n$  falling in histogram bin  $m$ , and  $E_m$  represents the energy at the midpoint of that bin. For the number of bins, typically, a value of  $M \approx 100$  can be used. Since Eq. (B.1.1) is linear in the  $\hat{Z}_k$ , the  $\hat{f}_k$  are unique up to an additive constant and we can fix one value, say  $f_1$ , by subtracting off the computed value of  $f_1^{(j+1)}$  after each iteration in order to avoid numerical drift.

After the initial guess has been reached, self-consistent iteration can rapidly refine the free energies to the desired tolerance while eliminating the bias arising from the use of histograms:

$$\hat{f}_i^{[j+1]} = -\ln \sum_{n=1}^N \left( \sum_{k=1}^K N_k \exp[\hat{f}_k^{[j]} - (\beta_k - \beta_i) E_n] \right)^{-1}$$

Again, we fix  $\hat{f}_1 = 0$  and terminate iterations when a relative tolerance

$$\rho = \max_{i=2,\dots,K} |\hat{f}_i^{[j+1]} - \hat{f}_i^{[j]}| / |\hat{f}_i^{[j]}|$$

is less than some given tolerance that ensures the computed expectations of properties of interest are no longer changing. We find that  $\rho < 10^{-7}$  is often a safe choice.

Cautions observed in Appendix C of [120] regarding sums of logarithms and numerical over/underflow in the evaluation of exponentials should be observed in implementation of this, or any, algorithm for obtaining the  $\hat{f}_i$ .

## B.2 PROOF THAT MODIFIED PT PROTOCOL SAMPLES FROM CANONICAL STATIONARY DISTRIBUTION

**Corollary 5.** *Here, we prove that the modified PT protocol described in Section 4.2.2 samples from the canonical stationary distribution at all temperatures.*

*Proof.* Define stationary distributions for momenta  $\mathbf{p}$  and coordinates  $\mathbf{q}$  in Cartesian space  $\mathbb{R}^{3N}$  at inverse temperature  $\beta$ :

$$\begin{aligned} \pi_p(\mathbf{p}|\beta) &= [P(\beta)]^{-1} e^{-\beta T(\mathbf{p})} \quad ; \quad P(\beta) = \int d\mathbf{p} \exp(-\beta T(\mathbf{p})) \\ \pi_q(\mathbf{q}|\beta) &= [Q(\beta)]^{-1} e^{-\beta U(\mathbf{q})} \quad ; \quad Q(\beta) = \int d\mathbf{q} \exp(-\beta U(\mathbf{q})) \end{aligned}$$

where  $T(\mathbf{p})$  denotes the kinetic energy and  $U(\mathbf{q})$  the potential energy function. Suppose we have two replicas whose current phase space points are denoted by  $\mathbf{z}_1 = (\mathbf{q}_1, \mathbf{p}_1)$  and  $\mathbf{z}_2 = (\mathbf{q}_2, \mathbf{p}_2)$ , initially at equilibrium at their respective inverse temperatures  $\beta_1$  and  $\beta_2$ , such that

$$\begin{aligned} \mathbf{p}_1 &\sim \pi_p(\mathbf{p}_1|\beta_1) \quad ; \quad \mathbf{q}_1 \sim \pi_q(\mathbf{q}_1|\beta_1) \\ \mathbf{p}_2 &\sim \pi_p(\mathbf{p}_2|\beta_2) \quad ; \quad \mathbf{q}_2 \sim \pi_q(\mathbf{q}_2|\beta_2). \end{aligned}$$

We now consider what happens to the distributions of  $\mathbf{z}_1$  and  $\mathbf{z}_2$  after an exchange attempt. Define ‘‘post-exchange attempt’’ coordinates and momenta for inverse temperature  $\beta_1$ :

$$\begin{aligned} \mathbf{q}'_1 &\leftarrow \begin{cases} \mathbf{q}_1 & \text{with prob. } 1 - \theta(\mathbf{q}_1, \mathbf{q}_2|\beta_1, \beta_2) \quad (\text{rejected}) \\ \mathbf{q}_2 & \text{with prob. } \theta(\mathbf{q}_1, \mathbf{q}_2|\beta_1, \beta_2) \quad (\text{accepted}) \end{cases} \\ \mathbf{p}'_1 &\sim \pi_p(\mathbf{p}'_1|\beta_1) \quad (\text{velocity randomization}) \end{aligned}$$

where the exchange acceptance probability  $\theta(\mathbf{q}_1, \mathbf{q}_2 | \beta_1, \beta_2)$  is given by

$$\theta(\mathbf{q}_1, \mathbf{q}_2 | \beta_1, \beta_2) = \min\{1, \exp[-\beta_1 U(\mathbf{q}_2) - \beta_2 U(\mathbf{q}_1) + \beta_1 U(\mathbf{q}_1) + \beta_2 U(\mathbf{q}_2)]\}$$

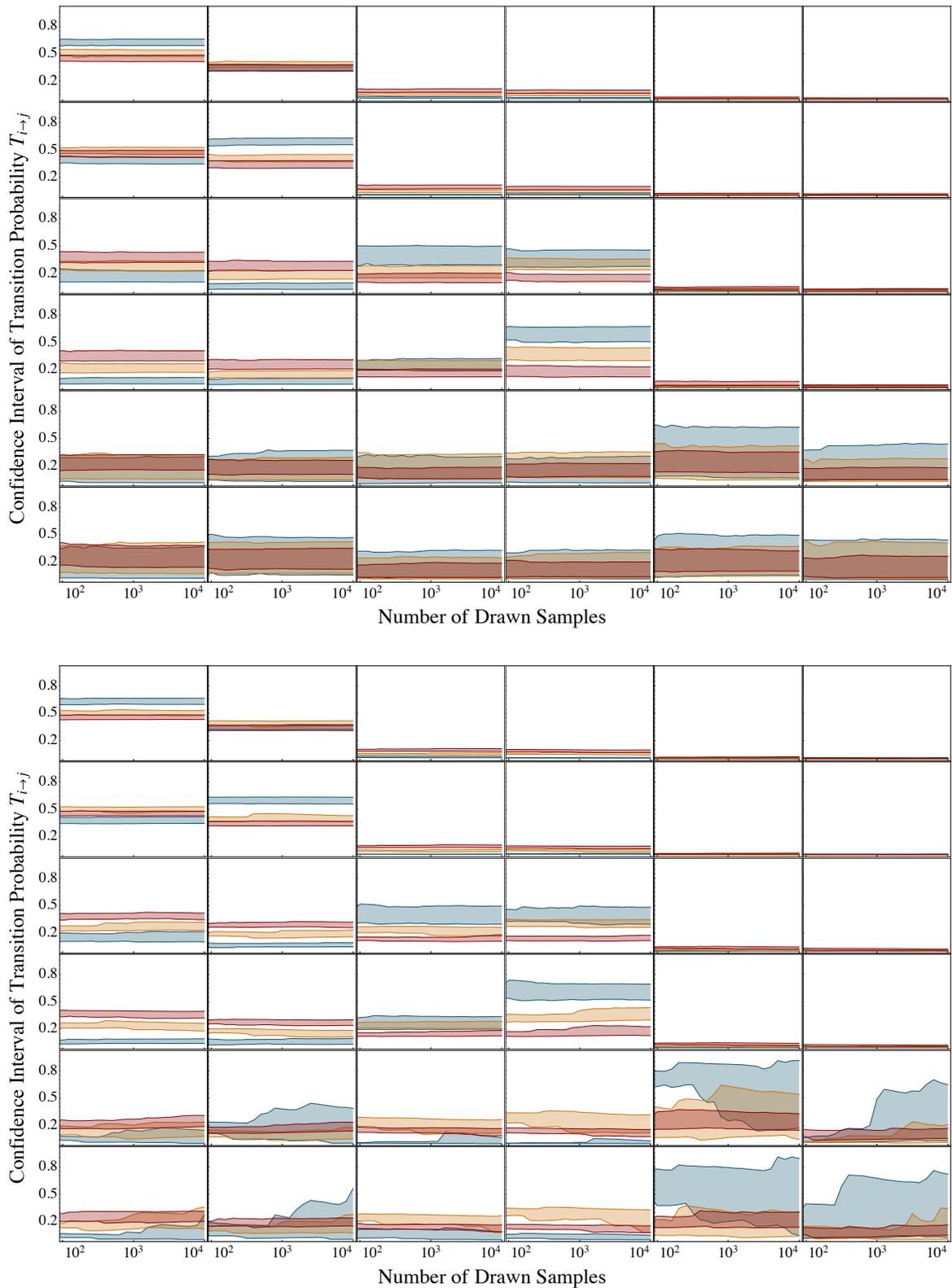
We now compute the distribution of  $\mathbf{q}'_1$ , the configuration supposedly at temperature  $\beta_1$  after the exchange attempt:

$$\begin{aligned} \rho_1(\mathbf{q}'_1) &= \int \mathbf{d}\mathbf{q}_2 [1 - \theta(\mathbf{q}'_1, \mathbf{q}_2 | \beta_1, \beta_2)] \pi_q(\mathbf{q}'_1 | \beta_1) \pi_q(\mathbf{q}_2 | \beta_2) \\ &\quad + \int \mathbf{d}\mathbf{q}_2 \theta(\mathbf{q}_2, \mathbf{q}'_1 | \beta_1, \beta_2) \pi_q(\mathbf{q}_2 | \beta_1) \pi_q(\mathbf{q}'_1 | \beta_2) \\ &= \int \mathbf{d}\mathbf{q}_2 [1 - \min\{1, e^{-\beta_1 U(\mathbf{q}_2)} e^{-\beta_2 U(\mathbf{q}'_1)} e^{+\beta_1 U(\mathbf{q}'_1)} e^{+\beta_2 U(\mathbf{q}_2)}\}] \times \\ &\quad \times \frac{e^{-\beta_1 U(\mathbf{q}'_1)}}{Q(\beta_1)} \frac{e^{-\beta_2 U(\mathbf{q}_2)}}{Q(\beta_2)} \\ &\quad + \int \mathbf{d}\mathbf{q}_2 \min\{1, e^{-\beta_1 U(\mathbf{q}'_1)} e^{-\beta_2 U(\mathbf{q}_2)} e^{+\beta_1 U(\mathbf{q}_2)} e^{+\beta_2 U(\mathbf{q}'_1)}\} \times \\ &\quad \times \frac{e^{-\beta_1 U(\mathbf{q}_2)}}{Q(\beta_1)} \frac{e^{-\beta_2 U(\mathbf{q}'_1)}}{Q(\beta_2)} \\ &= \frac{e^{-\beta_1 U(\mathbf{q}'_1)}}{Q(\beta_1)} \\ &\quad - \int \mathbf{d}\mathbf{q}_2 \min \left\{ \frac{e^{-\beta_1 U(\mathbf{q}'_1)}}{Q(\beta_1)} \frac{e^{-\beta_2 U(\mathbf{q}_2)}}{Q(\beta_2)}, \frac{e^{-\beta_1 U(\mathbf{q}_2)}}{Q(\beta_1)} \frac{e^{-\beta_2 U(\mathbf{q}'_1)}}{Q(\beta_2)} \right\} \\ &\quad + \int \mathbf{d}\mathbf{q}_2 \min \left\{ \frac{e^{-\beta_1 U(\mathbf{q}_2)}}{Q(\beta_1)} \frac{e^{-\beta_2 U(\mathbf{q}'_1)}}{Q(\beta_2)}, \frac{e^{-\beta_1 U(\mathbf{q}'_1)}}{Q(\beta_1)} \frac{e^{-\beta_2 U(\mathbf{q}_2)}}{Q(\beta_2)} \right\} \\ &= \pi_q(\mathbf{q}'_1 | \beta_1) \end{aligned}$$

Therefore, after the exchange attempt, the new configuration  $\mathbf{q}'_1$  is still at equilibrium at inverse temperature  $\beta_1$  (A similar series of steps can be applied for the temperature  $\beta_2$ ). Redrawing the momentum from the Maxwell-Boltzmann distribution at inverse temperature  $\beta_1$  will, of course, not change the equilibrium distribution, and can be shown to only support the canonical distribution at inverse temperature  $\beta_1$ , and no other stationary distribution [67]. Evolution by Hamiltonian dynamics for any length of time does not alter the stationary canonical distribution [159]. Therefore, the proposed protocol samples from the canonical distribution at the desired temperatures, provided sufficient time is allowed for equilibration.  $\square$

### B.3 CONVERGENCE OF TRANSITION PROBABILITIES IN BAYESIAN METHODS

The convergence of transition probabilities from the Bayesian sampling methods is presented in Figure B.1 for various temperatures.



**Figure B.1 – Confidence Levels of Transition Matrix Sampling** 95% confidence intervals of Transition probabilities sampled by the transition matrix estimation (upper plot) (TE) and rate matrix estimation (lower plot) (RE) versus number of drawn samples. Color indicates performance by temperature. Blue: 302 K, Yellow: 426 K, Red: 600 K. After about 5 000 samples the confidence intervals stabilize suggesting reasonably well sampled transition probabilities.

# C

---

## MATHEMATICAL DETAILS

---

### C.1 SENSITIVITY DERIVATIONS

Here we present the proofs to the lemmata 3 and 4:

**Lemma 3. Generalized Eigenvalue Sensitivity** *If the two (symmetric) observation correlation matrices  $\mathbf{A}(\theta) := \mathbf{C}(\tau_1; \theta)$  and  $\mathbf{B}(\theta) := \mathbf{C}(\tau_1; \theta)$  are depending on a scalar parameter  $\theta$  and  $\lambda$  and  $\mathbf{y}$  solve the generalized eigensystem problem*

$$\begin{aligned}(\mathbf{A} - \lambda \mathbf{B}) \mathbf{y} &= 0 \\ \mathbf{y}^T \mathbf{B} \mathbf{y} &= 1\end{aligned}$$

then the linear sensitivity  $\partial_\theta \lambda$  is related to  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\lambda$  and  $\mathbf{y}$  by

$$\frac{\partial \lambda}{\partial \theta} = \mathbf{y}^T \left( \frac{\partial \mathbf{A}}{\partial \theta} - \lambda \frac{\partial \mathbf{B}}{\partial \theta} \right) \mathbf{y}.$$

If  $\mathbf{A}$  or  $\mathbf{B}$  depend on more than one parameter these will just be added up. The case that  $\mathbf{A}$  and  $\mathbf{B}$  depend on different variables can be simplified by setting the appropriate derivatives to zero.

*Proof.* We differentiate the equation from the lemma by  $\theta$  to get

$$\frac{\partial \mathbf{A}}{\partial \theta} \mathbf{y} + \mathbf{A} \frac{\partial \mathbf{y}}{\partial \theta} - \lambda \mathbf{B} \frac{\partial \mathbf{y}}{\partial \theta} - \lambda \frac{\partial \mathbf{B}}{\partial \theta} \mathbf{y} - \frac{\partial \lambda}{\partial \theta} \mathbf{B} \mathbf{y} = 0$$

and collecting terms yields

$$(\mathbf{A} - \lambda \mathbf{B}) \frac{\partial \mathbf{y}}{\partial \theta} = - \left( \frac{\partial \mathbf{A}}{\partial \theta} - \lambda \frac{\partial \mathbf{B}}{\partial \theta} \right) \mathbf{y} + \frac{\partial \lambda}{\partial \theta} \mathbf{B} \mathbf{y}. \quad (\text{C.1.1})$$

Premultiplication with  $\mathbf{y}^T$  leads to

$$\mathbf{y}^T (\mathbf{A} - \lambda \mathbf{B}) \frac{\partial \mathbf{y}}{\partial \theta} = - \mathbf{y}^T \left( \frac{\partial \mathbf{A}}{\partial \theta} - \lambda \frac{\partial \mathbf{B}}{\partial \theta} \right) \mathbf{y} + \frac{\partial \lambda}{\partial \theta} \mathbf{y}^T \mathbf{B} \mathbf{y}.$$

and since we can rewrite the initial assumption as

$$\begin{aligned}(\mathbf{A} - \lambda \mathbf{B}) \mathbf{y} &= 0 \\ \Leftrightarrow \mathbf{y}^T (\mathbf{A}^T - \lambda \mathbf{B}^T) &= 0 \\ \Leftrightarrow \mathbf{y}^T (\mathbf{A} - \lambda \mathbf{B}) &= 0\end{aligned}$$

we can again simplify and the final solution takes the form

$$\frac{\partial \lambda}{\partial \theta} = \mathbf{y}^T \left( \frac{\partial \mathbf{A}}{\partial \theta} - \lambda \frac{\partial \mathbf{B}}{\partial \theta} \right) \mathbf{y}.$$

The extension to multiple dependent variables is obvious.  $\square$

We can in addition proof a similar lemma for the process vectors  $q_i$  which are related to the eigenvectors by a simple inversion. We write a simple lemma and proof

**Lemma 4. Process Vector Sensitivity** *If the two (symmetric) observation correlation matrices  $\mathbf{A}(\theta) := \mathbf{C}(\tau_1; \theta)$  and  $\mathbf{B}(\theta) := \mathbf{C}(\tau_1; \theta)$  are depending on a scalar parameter  $\theta$  and  $\lambda$  and  $\mathbf{y}$  solve the generalized eigensystem problem*

$$\begin{aligned} (\mathbf{A} - \lambda \mathbf{B}) \mathbf{y} &= 0 \\ \mathbf{y}^T \mathbf{B} \mathbf{y} &= 1 \end{aligned}$$

or more generally for the full set of generalized eigenvalues given by the diagonal matrix of generalized eigenvalues  $\mathbf{\Lambda}$  and the matrix of column-wise generalized eigenvectors  $\mathbf{Y}$

$$\begin{aligned} \mathbf{Y}^T (\mathbf{A} - \mathbf{\Lambda} \mathbf{B}) \mathbf{Y} &= 0 \\ \mathbf{Y}^T \mathbf{B} \mathbf{Y} &= Id \end{aligned} \tag{C.1.2}$$

then the linear sensitivity  $\partial_\theta \mathbf{y}$  is related to  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\lambda$  and  $\mathbf{y}$  by

$$\frac{\partial \lambda}{\partial \theta} = -(\mathbf{A} - \lambda \mathbf{B})^- \left( \frac{\partial \mathbf{A}}{\partial \theta} - \lambda \frac{\partial \mathbf{B}}{\partial \theta} \right) \mathbf{y} - \frac{1}{2} \left( \mathbf{y}^T \frac{\partial \mathbf{B}}{\partial \theta} \mathbf{y} \right) \mathbf{y}.$$

where

$$(\mathbf{A} - \lambda \mathbf{B})^- := \mathbf{Y} (\mathbf{\Lambda} - \lambda Id)^+ \mathbf{Y}^T$$

is the generalized symmetric inverse with  $(\ )^+$  representing the Moore-Penrose inverse which can in this case be written as

$$(\mathbf{\Lambda} - \lambda Id)^+ = \text{diag} \left( \begin{cases} \lambda^{-1} & \lambda_i \neq \lambda \\ 0 & \text{else} \end{cases} \right).$$

*Proof.* We start with the assumption that the generalized eigenvector  $\mathbf{y}$  is in column  $s$  and  $\lambda = \lambda_s$  and so  $\mathbf{Y}^{-1} \mathbf{y} = \mathbf{e}_s$  where  $\mathbf{e}_s$  is the canonical unit vector with zeros everywhere except at position  $s$ . From the definition of the generalized symmetric inverse it follows that

$$(\mathbf{\Lambda} - \lambda Id)^+ \mathbf{e}_s = (\lambda_s - \lambda) = 0$$

and we should note that the generalized symmetric inverse  $(\mathbf{A} - \lambda \mathbf{B})^-$  fulfills the first two Penrose condition (a weak Inverse of the multiplicative semigroup of matrices) but not the 3rd and 4th one,

stating that  $(\mathbf{A} - \lambda\mathbf{B})^-(\mathbf{A} - \lambda\mathbf{B})$  and  $(\mathbf{A} - \lambda\mathbf{B})(\mathbf{A} - \lambda\mathbf{B})^-$  are both hermitian. In fact, using

$$\mathbf{Y}^T\mathbf{B}\mathbf{Y} = \text{Id} \Leftrightarrow \mathbf{Y}^{-1} = \mathbf{Y}^T\mathbf{B} \Leftrightarrow \mathbf{B} = \mathbf{Y}^{-T}\mathbf{Y}^{-1}$$

we can compute

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{B})^-(\mathbf{A} - \lambda\mathbf{B}) &= (\mathbf{A} - \lambda\mathbf{B})(\mathbf{A} - \lambda\mathbf{B})^- \\ &= (\mathbf{A} - \lambda\mathbf{B})\mathbf{Y}(\mathbf{\Lambda} - \lambda\text{Id})^+\mathbf{Y}^T \\ &= \mathbf{Y}(\text{Id} - \mathbf{e}_s\mathbf{e}_s^T)\mathbf{Y}^{-1} \\ &= \text{Id} - \mathbf{y}\mathbf{y}^T\mathbf{B} \end{aligned}$$

and additionally show that

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{B})^-\mathbf{B}\mathbf{y} &= \mathbf{Y}(\mathbf{\Lambda} - \lambda\text{Id})\mathbf{Y}^T\mathbf{B}\mathbf{y} \\ &= \mathbf{Y}(\mathbf{\Lambda} - \lambda\text{Id})^+\mathbf{Y}^T\mathbf{Y}^{-T}\mathbf{Y}^{-1}\mathbf{y} \\ &= \mathbf{Y}(\mathbf{\Lambda} - \lambda\text{Id})^+\mathbf{Y}^{-1}\mathbf{y} \\ &= \mathbf{Y}(\mathbf{\Lambda} - \lambda\text{Id})\mathbf{e} \\ &= 0 \end{aligned} \tag{C.1.3}$$

Now reuse Eq. (C.1.1) from the previous lemma and premultiply both sides by  $(\mathbf{A} - \lambda\mathbf{A})^-$  to get

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{B})^-(\mathbf{A} - \lambda\mathbf{B})\frac{\partial\mathbf{y}}{\partial\theta} &= -(\mathbf{A} - \lambda\mathbf{B})^-\left(\frac{\partial\mathbf{A}}{\partial\theta} - \lambda\frac{\partial\mathbf{B}}{\partial\theta}\right)\mathbf{y} \\ &\quad + (\mathbf{A} - \lambda\mathbf{B})^-\mathbf{B}\mathbf{y}\frac{\partial\lambda}{\partial\theta} \end{aligned}$$

which can be simplified to

$$(\text{Id} - \mathbf{y}\mathbf{y}^T\mathbf{B})\frac{\partial\mathbf{y}}{\partial\theta} = -(\mathbf{A} - \lambda\mathbf{B})^-\left(\frac{\partial\mathbf{A}}{\partial\theta} - \lambda\frac{\partial\mathbf{B}}{\partial\theta}\right)\mathbf{y} \tag{C.1.4}$$

using that the last summand vanishes by Eq. (C.1.3). Differentiation of Eq. (C.1.2) then gives

$$\frac{\partial}{\partial\theta}(\mathbf{y}^T\mathbf{B}\mathbf{y}) = \mathbf{y}^T\frac{\partial\mathbf{B}}{\partial\theta}\mathbf{y} + 2\mathbf{y}^T\mathbf{B}\frac{\partial\mathbf{y}}{\partial\theta} = 0$$

which is equivalent to

$$\mathbf{y}^T\mathbf{B}\frac{\partial\mathbf{y}}{\partial\theta} = \frac{1}{2}\mathbf{y}^T\frac{\partial\mathbf{B}}{\partial\theta}\mathbf{y} \tag{C.1.5}$$

We use Eq. (C.1.4) in Eq. (C.1.5) and finally get

$$\begin{aligned} (\text{Id} - \mathbf{y}\mathbf{y}^T\mathbf{B})\frac{\partial\mathbf{y}}{\partial\theta} &= -(\mathbf{A} - \lambda\mathbf{B})^-\left(\frac{\partial\mathbf{A}}{\partial\theta} - \lambda\frac{\partial\mathbf{B}}{\partial\theta}\right)\mathbf{y} \\ \frac{\partial\mathbf{y}}{\partial\theta} &= -(\mathbf{A} - \lambda\mathbf{B})^-\left(\frac{\partial\mathbf{A}}{\partial\theta} - \lambda\frac{\partial\mathbf{B}}{\partial\theta}\right)\mathbf{y} \\ &\quad + \frac{1}{2}\left(\mathbf{y}^T\frac{\partial\mathbf{B}}{\partial\theta}\mathbf{y}\right)\mathbf{y} \end{aligned}$$

which proves the lemma.  $\square$

## C.2 MOORE-PENROSE PSEUDOINVERSE

Although the Moore-Penrose pseudoinverse can be defined for arbitrary matrices over the complex numbers, we are only interested in real-valued matrices  $\mathbf{A} \in \mathbb{R}^{n \times m}$ . The Moore-Penrose pseudoinverse  $[\mathbf{A}]^+$  of  $\mathbf{A}$  is then defined as matrix fulfilling the following four properties:

1.  $\mathbf{A} [\mathbf{A}]^+ \mathbf{A} = \mathbf{A}$
2.  $[\mathbf{A}]^+ \mathbf{A} [\mathbf{A}]^+ = [\mathbf{A}]^+$
3.  $[\mathbf{A}]^+ \mathbf{A} = ([\mathbf{A}]^+ \mathbf{A})^T$
4.  $\mathbf{A} [\mathbf{A}]^+ = (\mathbf{A} [\mathbf{A}]^+)^T$

This pseudoinverse exists, is unique and can be expressed using a Singular Value Decomposition (SVD) of

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

by

$$[\mathbf{\Sigma}]^+ = \mathbf{V} [\mathbf{\Sigma}]^+ \mathbf{U}^T$$

where the pseudoinverse of the diagonal matrix  $\mathbf{\Sigma}$  is given by the inverse diagonal element which are non zero:

$$[\mathbf{\Sigma}]^+ = \text{diag} \left( \begin{cases} \sigma_i^{-1} & \sigma_i \neq 0 \\ 0 & \sigma_i = 0 \end{cases} \right).$$

Thus, the pseudoinverse can be regarded as an inverse acting only on the image space.

## C.3 RATE ESTIMATION ERROR BOUNDS

This section contains the derivations of the bounds for single and multi- $\tau$  estimations using in chapter 5.

C.3.1 Single- $\tau$  rate estimator

We start with the autocorrelation function of the observable  $\zeta(x)$  used to estimate the dominant rate  $\hat{\kappa}_2$  by

$$\begin{aligned} \hat{\kappa}_2 &= -\tau^{-1} \ln \tilde{\lambda}_2(\tau) \\ &= -\tau^{-1} \ln \langle \zeta(x_t) \zeta(x_{t+\tau}) \rangle \end{aligned}$$

and make a spectral decomposition that can be written as

$$\hat{\kappa}_2 = -\tau^{-1} \ln \left( \alpha \lambda_2(\tau) + \sum_{i>2} q_i^2 \lambda_i(\tau) \right)$$

or in terms of the exact rates  $\kappa_i$  by

$$\hat{\kappa}_2 = -\tau^{-1} \ln \left( \alpha e^{-\tau\kappa_2} + \sum_{i>2} q_i^2 e^{-\tau\kappa_i} \right).$$

Separation of the original dominant timescale  $\kappa_2$ , yields

$$\begin{aligned} \hat{\kappa}_2 &= -\tau^{-1} \ln \left( e^{-\tau\kappa_2} \left( \alpha + \sum_{i>2} q_i^2 e^{-\tau(\kappa_i - \kappa_2)} \right) \right) \\ &= -\tau^{-1} \left( \ln e^{-\tau\kappa_2} + \ln \left( \alpha + \sum_{i>2} q_i^2 e^{-\tau(\kappa_i - \kappa_2)} \right) \right) \\ &= \kappa_2 - \tau^{-1} \ln \left( \alpha + \sum_{i>2} q_i^2 e^{-\tau(\kappa_i - \kappa_2)} \right) \end{aligned}$$

and we get

$$\begin{aligned} \hat{\kappa}_2 - \kappa_2 = \Delta\kappa_2 &= -\tau^{-1} \ln \left( \alpha + \sum_{i>2} q_i^2 e^{-\tau(\kappa_i - \kappa_2)} \right) \\ &= \tau^{-1} \ln(\alpha^{-1}) - \ln \left( 1 + \sum_{i>2} \frac{q_i^2}{\alpha} e^{-\tau(\kappa_i - \kappa_2)} \right) \end{aligned}$$

as an expression for the estimation error. Since the second term is strictly positive this gives

$$\Delta\kappa_2 \leq \tau^{-1} \ln \alpha^{-1}$$

as an upper bound so that the error is in general dominated by a  $\tau^{-1}$ -dependence. Using  $\kappa_i > \kappa_2$  for  $i > 2$  and  $\ln(1+x) < x$  we can also find a lower bound expressed by

$$\begin{aligned} \Delta\kappa_2 &> \tau^{-1} \ln \alpha^{-1} - \alpha^{-1} \ln \left( 1 + \sum_{i>2} (q_i^2 / \alpha) e^{-\tau(\kappa_i - \kappa_2)} \right) \\ &> \tau^{-1} \ln \alpha^{-1} - \alpha^{-1} \ln \left( 1 + \frac{1 - \alpha}{\alpha} e^{-\tau(\kappa_3 - \kappa_2)} \right) \\ &> \tau^{-1} \ln \alpha^{-1} - \alpha^{-1} \frac{1 - \alpha}{\alpha} e^{-\tau(\kappa_3 - \kappa_2)}. \end{aligned}$$

In the two-state case with  $\kappa_i \gg \kappa_2$ ,  $i > 2$  we find that

$$\Delta\kappa_2 \lesssim \tau^{-1} \ln \alpha^{-1}$$

### c.3.2 Multi- $\tau$ exponential fitting estimators

We will calculate the systematic error of estimating  $\kappa_2$  via an exponential fit to the function:

$$\begin{aligned} \tilde{\lambda}_2 = \langle \tilde{\xi}(x_t) \tilde{\xi}(x_{t+\tau}) \rangle &= \langle \psi_2, \tilde{\xi} \rangle e^{-\tau\kappa_2} + \sum_{i>2} \langle \psi_i, \tilde{\xi} \rangle e^{-\tau\kappa_i} \\ &= \alpha e^{-\tau\kappa_2} + \sum_{i>2} q_i^2 e^{-\tau\kappa_i} \end{aligned}$$

using a set of  $m$  lagtimes  $\tau_i \in \{\tau_1, \dots, \tau_m\}$ . This fitting procedure can for example be done by a least-squares linear fit to the logarithmized data points

$$\begin{aligned} z(\tau) &= \ln\langle \zeta(x_t)\zeta(x_{t+\tau}) \rangle \\ &= \ln\left( \alpha e^{-\tau\kappa_2} + \sum_{i>2} q_i^2 e^{-\tau\kappa_i} \right). \end{aligned}$$

A simple linear regression algorithm can be expressed as

$$\begin{aligned} \hat{\kappa}_2 &= \frac{\mathbb{E}[\tau_j]\mathbb{E}[z(\tau_j)] - \mathbb{E}[\tau_j z(\tau_j)]}{\text{Var}(\tau)} \\ &= \frac{\mathbb{E}[(\mathbb{E}[\tau_j] - \tau_j)z(\tau_j)]}{\text{Var}(\tau)} \\ &= \frac{m^{-1} \sum_{j=1}^m (\bar{\tau} - \tau_j)z(\tau_j)}{\text{Var}(\tau)} \end{aligned}$$

with

$$\bar{\tau} = \mathbb{E}[\tau_j] = m^{-1} \sum_{j=1}^m \tau_j$$

and

$$\text{Var}(\tau) = \mathbb{E}[(\tau_j - \bar{\tau})^2] = m^{-1} \sum_{j=1}^m \tau_j^2 - \bar{\tau}^2.$$

Using the same transformations as in the single- $\tau$  case we get

$$\Delta\kappa_2 = m^{-1} \sum_{j=1}^m \frac{(\bar{\tau} - \tau_j)}{\text{Var}(\tau)} \ln\left( 1 + \sum_{i>2} (q_i^2/\alpha) e^{-\tau_j(\kappa_i - \kappa_2)} \right)$$

for the estimation error for this particular scheme. This expression can be bounded from above by keeping only all positive summands  $\tau_j < \bar{\tau}$ ,  $j \leq m^+ \leq m - 1$ , using the increasing ordering of lag times  $\tau_j$ , to get

$$0 < \Delta\kappa_2 < \sum_{j=1}^{m^+} \frac{(\bar{\tau} - \tau_j)}{m \text{Var}(\tau)} \ln\left( 1 + \sum_{i>2} (q_i^2/\alpha) e^{-\tau_j(\kappa_i - \kappa_2)} \right),$$

then replacing all summands by the largest one  $j = 1$  with the result

$$\begin{aligned} 0 < \Delta\kappa_2 &< \frac{m^+ (\bar{\tau} - \tau_1)}{m \text{Var}(\tau)} \ln\left( 1 + \sum_{i>2} (q_i^2/\alpha) e^{-\tau_1(\kappa_i - \kappa_2)} \right) \\ &< \frac{m - 1 (\bar{\tau} - \tau_1)}{m \text{Var}(\tau)} \ln\left( 1 + \sum_{i>2} (q_i^2/\alpha) e^{-\tau_1(\kappa_i - \kappa_2)} \right) \end{aligned}$$

where we bounded the number of positive summands  $m^+$  by  $m - 1$ . Using again that the rates  $\kappa_i > \kappa_2$  are ordered we can derive a bound that only depends on the gap  $\kappa_3 - \kappa_2$  and the chosen lag times  $\tau_i$

$$0 < \Delta\kappa_2 < \frac{m-1}{m} \frac{(\bar{\tau} - \tau_1)}{\text{Var}(\tau)} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right).$$

Lastly, using Samuelson's inequality and  $(a+b)^2 \geq 2ab$  we can bound this by

$$0 < \Delta\kappa_2 < \frac{(m-1)^2}{m(\tau_m - \bar{\tau})} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right)$$

without knowledge about the variance. It is important to note that the error is now dominated by an exponential decay in the gap between 2nd and 3rd timescale  $\kappa_3 - \kappa_2$  and also decays exponentially in the minimal lagtime  $\tau_1$ . For  $\alpha \rightarrow 1$  the error vanishes as it must be.

In the special case of two lagtimes  $\tau_1$  and  $\tau_2$  we get

$$\begin{aligned} 0 < \Delta\kappa_2 &< \frac{1}{2} \frac{(1/2(\tau_1 + \tau_2) - \tau_1)}{1/4(\tau_2 + \tau_2)^2} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right) \\ &< (\tau_2 - \tau_1)^{-1} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right) \\ &< (\tau_2 - \tau_1)^{-1} \frac{1-\alpha}{\alpha} \left( e^{-\tau_1(\kappa_3 - \kappa_2)} \right) \end{aligned}$$

and in particular for  $\tau_2 = 2\tau_1 = 2\tau$  this simplifies to

$$0 < \Delta\kappa_2 \lesssim \tau^{-1} \frac{1-\alpha}{\alpha} \left( e^{-\tau(\kappa_3 - \kappa_2)} \right)$$

where we now have an exponential decay in  $\tau$  compared to the  $\tau^{-1}$ -dependence in the single- $\tau$  estimation case. For the case of  $m$  lagtimes that are spaced equidistantly  $\tau_k = k \cdot \tau$  we derive similarly

$$\begin{aligned} 0 < \Delta\kappa_2 &< \frac{m-1}{m} \frac{(1/2(m-1) \cdot \tau)}{1/12(m^2-1) \cdot \tau^2} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right) \\ &< \frac{m-1}{m} \frac{1}{1/6(m+1) \cdot \tau} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right) \\ &< \text{const} \cdot \tau^{-1} \ln \left( 1 + e^{-\tau_1(\kappa_3 - \kappa_2)} \frac{1-\alpha}{\alpha} \right) \\ &< \text{const} \cdot \tau^{-1} \frac{1-\alpha}{\alpha} \left( e^{-\tau_1(\kappa_3 - \kappa_2)} \right) \end{aligned}$$

where the constant is  $\mathcal{O}(m^{-1})$ . Finally, the case of a single- $\tau$  estimation can be recovered by the choice of  $\tau_1 = 0$  and  $\tau_2 = \tau$ , in which case we get

$$\begin{aligned} 0 < \Delta\kappa_2 &< \tau^{-1} \ln(1 + ((1-\alpha)/\alpha)) \\ &< \tau^{-1} \ln \frac{1}{\alpha}. \end{aligned}$$

Note, that in the absence of statistical uncertainty it is always better to estimate using the two largest lagtimes  $\tau_{m-1}$  and  $\tau_m$  which we do not prove here.



# D

---

## SYSTEMS SETUP

---

### D.1 EXEMPLARY MODEL SYSTEMS

The model systems used to illustrate concepts and methods are chosen to mimic a simple diffusion in a potential and are set up in the following manner: First, a  $d$ -dimensional potential function ( $d \in \{1, 2, 3\}$ ) in units of  $k_B T$  is defined as  $V(\mathbf{x})$  with  $\mathbf{x} \in \mathbb{R}^d$ . This potential was then evaluated on an equidistant lattice of  $N^d$  lattice points in a selected range  $[x^{\min}, x^{\max}]^d$ . For reasons of simplicity all models for  $d > 1$  are kept quadratic or cubic and the lattice points are indexed by an integer vector  $\mathbf{k} \in \mathbb{N} = \{1, \dots, N\}^d$ . On this lattice, a simple jump process was defined, where proposed jumps are only allowed between direct neighbors

$$\mathbb{P}^{\text{jump}}(\mathbf{i} \rightarrow \mathbf{j}) = \frac{1}{n_{\mathbf{i}}} \begin{cases} 1 & \text{if } |\mathbf{i} - \mathbf{j}| = 1 \\ 0 & \text{else} \end{cases},$$

to mimic the diffusion, e.g. from  $(x)$  to  $\{(x-1), (x+1)\}$  in the 1-dimensional case. The normalization  $n_{\mathbf{i}}$  equals the total number of neighbors of state  $\mathbf{i}$  so that

$$\sum_{\mathbf{j} \in \mathbb{N}} \mathbb{P}^{\text{jump}}(\mathbf{i} \rightarrow \mathbf{j}) = 1, \forall \mathbf{i}$$

holds. For the construction of a transition matrix, the tuples  $\mathbf{k}$  are mapped to a single integer (state) using the one-to-one and onto mapping  $D$ ,

$$\begin{aligned} D &: \mathbf{k} = \{k_1, \dots, k_d\} \in \{1, \dots, N\}^d \\ &\mapsto D(\mathbf{k}) = 1 + \sum_{i=0}^{d-1} N^i \cdot (k_i - 1) \in \{1, \dots, N^d\} \end{aligned}$$

and its inverse  $D^{-1}$ ,

$$\begin{aligned} D^{-1} &: n \in \{1, \dots, N^d\} \mapsto D^{-1}(n)_k \\ &= 1 + \left( \left\lfloor (n-1)/N^{(k-1)} \right\rfloor \bmod N \right) \end{aligned}$$

where we used  $V_i \equiv V(D^{-1}(i))$  and  $\lfloor \cdot \rfloor$  indicating the floor function. This allows to construct a discrete state transition matrix

$$\mathbf{T} \in [0, 1]^{N^d \times N^d}$$

and we choose the transition probabilities between the current state  $i = D(\mathbf{i})$  and a target state  $j = D(\mathbf{j})$  to be

$$T_{ij} = \mathbb{P}(D^{-1}(i) \rightarrow D^{-1}(j)) \times \\ \times \min\{1, \exp(-(V_i - V_j)) \frac{\mathbb{P}^{\text{jump}}(D^{-1}(j) \rightarrow D^{-1}(i))}{\mathbb{P}^{\text{jump}}(D^{-1}(i) \rightarrow D^{-1}(j))}\}, i \neq j$$

and for the diagonal elements

$$T_{ii} = 1 - \sum_{j \neq i} T_{ij}$$

in accordance with the Metropolis-Hastings algorithm[160]. This ensures the correct transition probabilities also in the case of non-symmetric proposal steps that happen in the boundary area and so defines a reversible Markov process which has a stationary distribution  $\pi$  given by the Boltzmann distribution

$$\pi_i \propto \exp(-V_i).$$

Since the dynamics is Markovian by definition, all Markov model calculations can be executed exactly, providing an unambiguous reference for our analysis of Markov models. In order to model metastable potentials, we often use symmetric gaussian potential basins defined by

$$B[I, \boldsymbol{\mu}, \sigma^2](\mathbf{x}) = I_k \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})}{2\sigma^2}\right).$$

The potentials used in this thesis are then defined as follows:

1. Diffusion in a **one-dimensional four-well potential** with  $N = 100$  discretization points in the range  $[-1, 1]$  (used in chapters 2 and 5)

$$V(x) = 4 \left( x^8 + 0.8e^{-80x^2} + 0.2e^{-80(x-0.5)^2} + 0.5e^{-40(x+0.5)^2} \right)$$

2. Diffusion in a **two-dimensional three-well potential** with  $N^2 = 30 \times 30$  discretization points in the range  $[1, 30]^2$  (used in chapters 2 and 3)

$$V(\mathbf{x}) = B[2, \{15, 15\}, 200](\mathbf{x}) - B[1.2, \{9, 9\}, 12.5](\mathbf{x}) \\ - B[0.8, \{21, 9\}, 12.5](\mathbf{x}) - B[1.0, \{13, 21\}, 12.5](\mathbf{x})$$

3. Diffusion in a **three-dimensional 5-well potential** with  $N^3 = 100 \times 100 \times 100$  discretization points in the range  $[-1, 1]^2$  (used in chapter 3)

$$V(\mathbf{x}) = \sum B[b_i / \sqrt{2\pi\sigma_i^2}, \boldsymbol{\mu}_i, \sigma_i^2](\mathbf{x})$$

with parameters

$i$	Sign $b_i$	Mean $\mu_i$	Variance $\sigma_i^2$
1	-1	$\{0.0, 0.0, -0.2\}$	$0.10^2$
2	-1	$\{-0.6, 0.2, -0.6\}$	$0.08^2$
3	-1	$\{-0.6, 0.4, 0.4\}$	$0.08^2$
4	+1	$\{0.4, -0.6, -0.6\}$	$0.05^2$
5	-1	$\{-0.6, -0.6, -0.6\}$	$0.05^2$

4. A diffusion in a **two-dimensional 2-state model symmetric in the  $x_2$ -axis** without actual discretization in the range  $[-4, 4]^2$  (used in chapter 5)

$$V(\mathbf{x}) = 0.25x_1^4 - 2.5x_1^2 + 0.5x_1 + 0.5x_2^2.$$

This system is only analyzed on 1-dimensional projections at 3 different viewing angles  $\alpha = \{0^\circ, 45^\circ, 72^\circ\}$ , where the angle is chosen with respect to the  $x_1$  coordinate providing the best possible projection angle at  $\alpha = 0^\circ$ . In these three cases a discretization with  $N = 40$  discretization points in the range  $[-4, 4]$  is used.

## D.2 RNA HAIRPIN

For the demonstration of the spectral estimation procedure in chapter 5 the method was applied to measurements of a single p5ab RNA hairpin in an optical trap observed under passive conditions (see Figure 5.7 for the hairpin structure). We would like to acknowledge Philip Elms who conducted the data generating experiments at the Marqusee lab at UC Berkeley the results of which are submitted for publication [147]. For this reason we will refer to the information given in the arxiv-preprint in Ref. [133] in the following:

Under the influence of an external biasing (pulling) force, the hairpin exhibits an apparent two-state kinetics in the induced transition from the unfolded (extended) to the folded (compact) state and has been studied previously in single-molecule force spectroscopy experiments [161, 162, 147]. Recorded was the instantaneous force on the optically trapped bead along the bead-bead axis, which was then converted to a total bead-to-bead extension using the theory described in section 5.10.

In the experiment a total of 10 hairpins were observed. For a single trapped hairpin up to 25 trajectories of 60s length were recorded at various total trap extensions using a dual-beam counter-propagating optical trap [163, 164]. The trap extensions were chosen to go from

completely unfolded states to only folded and back provided that the hairpin *survived* the complete measuring procedure. The high sampling rate of 50 kHz lead to a total of 3,000,000 data points per trace and was far above the corner frequency for bead response under the experimental conditions. The data was also subsampled to a frequency of 1 kHz, which was below the corner frequency of the bead, so that the bead velocity has decorrelated between sequential observations due to hydrodynamic interactions [147].

For the present analysis hairpin no. 3 was selected and only the first 13 traces of the *transition* from folded to unfolded were taken into consideration. Out of this set only 6 traces were analyzed: Chosen were trace no. 2 (completely folded) and no. 13 (completely unfolded) as extreme cases for comparison as well as all traces that showed a clear 2-state histogram, trace no. 7-11. Finally trace no. 9 was also discarded for clarity since it was very similar to trace no. 8 and provided no further insight into the dynamics. See Figure 5.5 for a plot of the traces along with their histograms.

---

## BIBLIOGRAPHY

---

- [1] Prinz, J.-H., Wu, H., Sarich, M., Keller, B. G., Senne, M., Held, M., Chodera, J. D., Schütte, C. & Noé, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011). doi [10.1063/1.3565032](https://doi.org/10.1063/1.3565032). (Cited on pages [ix](#), [4](#), [7](#), [34](#), [40](#), [44](#), [47](#), [98](#), [99](#), [112](#), and [114](#).)
- [2] Prinz, J.-H., Held, M., Smith, J. C. & Noé, F. Efficient Computation, Sensitivity, and Error Analysis of Committor Probabilities for Complex Dynamical Processes. *Multiscale Model. Sim.* **9**, 545–567 (2011). doi [10.1137/100789191](https://doi.org/10.1137/100789191). (Cited on pages [ix](#), [4](#), [41](#), [42](#), and [49](#).)
- [3] Prinz, J.-H., Chodera, J. D., Pande, V. S., Swope, W. C., Smith, J. C. & Noé, F. Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics. *J. Chem. Phys.* **134**, 244108 (2011). doi [10.1063/1.3592153](https://doi.org/10.1063/1.3592153). (Cited on pages [ix](#), [77](#), and [143](#).)
- [4] Chodera, J. D., Swope, W. C., Noé, F., Prinz, J.-H., Shirts, M. R. & Pande, V. S. Dynamical reweighting: improved estimates of dynamical properties from simulations at multiple temperatures. *J. Chem. Phys.* **134**, 244107 (2011). doi [10.1063/1.3592152](https://doi.org/10.1063/1.3592152). (Cited on pages [ix](#), [77](#), [78](#), [79](#), [80](#), [81](#), [82](#), [84](#), [94](#), and [143](#).)
- [5] Prinz, J.-H., Keller, B. G. & Noé, F. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.* **13**, 16912–16927 (2011). doi [10.1039/c1cp21258c](https://doi.org/10.1039/c1cp21258c). (Cited on page [ix](#).)
- [6] Keller, B. G., Prinz, J.-H. & Noé, F. Markov models and dynamical fingerprints: Unraveling the complexity of molecular kinetics. *Chemical Physics* **396**, 92–107 (2011). doi [10.1016/j.chemphys.2011.08.021](https://doi.org/10.1016/j.chemphys.2011.08.021). (Cited on pages [ix](#) and [104](#).)
- [7] Held, M., Metzner, P., Prinz, J.-H. & Noé, F. Mechanisms of protein-ligand association and its modulation by protein mutations. *Biophys. J.* **100**, 701–710 (2011). doi [10.1016/j.bpj.2010.12.3699](https://doi.org/10.1016/j.bpj.2010.12.3699). (Cited on pages [ix](#) and [75](#).)
- [8] Prinz, J.-H., Chodera, J. D. & Noé, F. Spectral rate theory for two-state kinetics. *submitted to Physical Review X, ArXiv preprint physics* (2012). url <http://arxiv.org/abs/1207.0225>. (Cited on pages [ix](#) and [97](#).)

- [9] Jäger, M., Zhang, Y., Bieschke, J., Nguyen, H., Dendle, M., Bowman, M. E., Noel, J. P., Gruebele, M. & Kelly, J. W. Structure-function-folding relationship in a WW domain. *Proc. Nat. Acad. Sci. USA* **103**, 10648–10653 (2006). doi [10.1073/pnas.0600511103](https://doi.org/10.1073/pnas.0600511103). (Cited on page 1.)
- [10] Kobitski, A. Y., Nierth, A., Helm, M., Jaschke, A. & Nienhaus, U. G. Mg<sup>2+</sup>-dependent folding of a Diels-Alderase ribozyme probed by single-molecule FRET analysis. *Nuc. Acid. Res.* **35**, 2047–2059 (2007). doi [10.1093/nar/gkm072](https://doi.org/10.1093/nar/gkm072). (Cited on page 1.)
- [11] Noé, F., Krachtus, D., Smith, J. C. & Fischer, S. Transition networks for the comprehensive characterization of complex conformational change in proteins. *J. Chem. Theo. Comp.* **2**, 840–857 (2006). doi [10.1021/ct050162r](https://doi.org/10.1021/ct050162r). (Cited on pages 1 and 3.)
- [12] Fischer, S., Windshügel, B., Horak, D., Holmes, K. C. & Smith, J. C. Structural mechanism of the recovery stroke in the myosin molecular motor. *Proc. Nat. Acad. Sci. USA* **102**, 6873–6878 (2005). doi [10.1073/pnas.0408784102](https://doi.org/10.1073/pnas.0408784102). (Cited on page 1.)
- [13] Ostermann, A., Waschipky, R., Parak, F. G. & Nienhaus, G. U. Ligand binding and conformational motions in myoglobin. *Nature* **404**, 205–208 (2000). doi [10.1038/35004622](https://doi.org/10.1038/35004622). (Cited on page 1.)
- [14] Frauenfelder, H., Sligar, S. & Wolynes, P. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991). doi [10.1126/science.1749933](https://doi.org/10.1126/science.1749933). (Cited on page 1.)
- [15] Wensley, B. G., Batey, S., Bone, F. A. C., Chan, Z. M., Tumelty, N. R., Steward, A., Kwa, L. G., Borgia, A. & Clarke, J. Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature* **463**, 685–688 (2010). doi [10.1038/nature08743](https://doi.org/10.1038/nature08743). (Cited on pages 1 and 114.)
- [16] Gebhardt, J. C. M., Bornschlöggl, T. & Rief, M. Full distance-resolved folding energy landscape of one single protein molecule. *Proc. Nat. Acad. Sci. USA* **107**, 2013–2018 (2010). doi [10.1073/pnas.0909854107](https://doi.org/10.1073/pnas.0909854107). (Cited on pages 1 and 114.)
- [17] Santoso, Y., Joyce, C. M., Potapova, O., Le Reste, L., Hohlbein, J., Torella, J. P., Grindley, N. D. F. & Kapanidis, A. N. Conformational transitions in DNA polymerase I revealed by single-molecule FRET. *Proc. Nat. Acad. Sci. USA* **107**, 715–720 (2010). doi [10.1073/pnas.0910909107](https://doi.org/10.1073/pnas.0910909107). (Cited on pages 1 and 114.)
- [18] Neubauer, H. *et al.* Orientational and dynamical heterogeneity of rhodamine 6G terminally attached to a DNA helix revealed by NMR and single-molecule fluorescence spectroscopy. *J. Am.*

- Chem. Soc.* **129**, 12746–12755 (2007). doi [10.1021/ja0722574](https://doi.org/10.1021/ja0722574). (Cited on pages [1](#) and [114](#).)
- [19] Gansen, A., Valeri, A., Hauger, F., Felekyan, S., Kalinin, S., Tóth, K., Langowski, J. & Seidel, C. A. M. Nucleosome disassembly intermediates characterized by single-molecule FRET. *Proc. Nat. Acad. Sci. USA* **106**, 15308–15313 (2009). doi [10.1073/pnas.0903005106](https://doi.org/10.1073/pnas.0903005106). (Cited on pages [1](#) and [114](#).)
- [20] Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E. & Kern, D. Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117–121 (2005). doi [10.1038/nature04105](https://doi.org/10.1038/nature04105). (Cited on pages [1](#) and [114](#).)
- [21] Min, W., Luo, G., Cherayil, B. J., Kou, S. C. & Xie, X. S. Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Phys. Rev. Lett* **94**, 198302 (2005). doi [10.1103/PhysRevLett.94.198302](https://doi.org/10.1103/PhysRevLett.94.198302). (Cited on pages [1](#) and [114](#).)
- [22] English, B. P., Min, W., van Oijen, A. M., Lee, K. T., Luo, G., Sun, H., Cherayil, B. J., Kou, S. C. & Xie, X. S. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol.* **2**, 87–94 (2006). doi [10.1038/nchembio759](https://doi.org/10.1038/nchembio759). (Cited on page [1](#).)
- [23] Waldauer, S. A. *et al.* Ruggedness in the folding landscape of protein L. *HFSP Journal* **2**, 388–395 (2008). doi [10.2976/1.3013702](https://doi.org/10.2976/1.3013702). (Cited on page [1](#).)
- [24] Mello, C. C. & Barrick, D. An experimentally determined protein folding energy landscape. *Proc. Nat. Acad. Sci. USA* **101**, 14102–14107 (2004). doi [10.1073/pnas.0403386101](https://doi.org/10.1073/pnas.0403386101). (Cited on page [1](#).)
- [25] Matagne, A., Radford, S. E. & Dobson, C. M. Fast and slow tracks in lysozyme folding: insight into the role of domains in the folding process. *J. Mol. Biol.* **267**, 1068–1074 (1997). doi [10.1006/jmbi.1997.0963](https://doi.org/10.1006/jmbi.1997.0963). (Cited on page [1](#).)
- [26] Goldbeck, R., Thomas, Y., Chen, E., Esquerra, R. & Kliger, D. Multiple pathways on a protein-folding energy landscape: Kinetic evidence. *Proc. Nat. Acad. Sci. USA* **96**, 2782–2787 (1999). doi [10.1073/pnas.96.6.2782](https://doi.org/10.1073/pnas.96.6.2782). (Cited on page [1](#).)
- [27] Sridevi, K., Juneja, J., Bhuyan, A. K., Krishnamoorthy, G. & Udgaonkar, J. B. The slow folding reaction of barstar: the core tryptophan region attains tight packing before substantial secondary and tertiary structure formation and final compaction of the polypeptide chain. *J. Mol. Biol.* **302**, 479–495 (2000). doi [10.1006/jmbi.2000.4060](https://doi.org/10.1006/jmbi.2000.4060). (Cited on page [1](#).)

- [28] Lindberg, M. O. & Oliveberg, M. Malleability of protein folding pathways: a simple reason for complex behaviour. *Curr. Opin. Struct. Biol.* **17**, 21–29 (2007). doi [10.1016/j.sbi.2007.01.008](https://doi.org/10.1016/j.sbi.2007.01.008). (Cited on page [1](#).)
- [29] Eisenmesser, E. Z., Bosco, D. A., Akke, M. & Kern, D. Enzyme dynamics during catalysis. *Science* **295**, 1520–1523 (2002). doi [10.1126/science.1066176](https://doi.org/10.1126/science.1066176). (Cited on page [1](#).)
- [30] Feng, H., Zhou, Z. & Bai, Y. S. A protein folding pathway with multiple folding intermediates at atomic resolution. *Proc. Nat. Acad. Sci. USA* **102**, 5026–5031 (2005). doi [10.1073/pnas.0501372102](https://doi.org/10.1073/pnas.0501372102). (Cited on page [1](#).)
- [31] Smith, G., Lee, K., Qu, X., Xie, Z., Pesic, J., Sosnick, T., Pan, T. & Scherer, N. A large collapsed-state RNA can exhibit simple exponential single-molecule dynamics. *J. Mol. Biol.* **378**, 941–951 (2008). doi [10.1016/j.jmb.2008.01.078](https://doi.org/10.1016/j.jmb.2008.01.078). (Cited on page [1](#).)
- [32] Zhuang, X. & Rief, M. Single-molecule folding. *Curr. Opin. Struct. Biol.* **13**, 88–97 (2003). doi [10.1016/S0959-440X\(03\)00011-3](https://doi.org/10.1016/S0959-440X(03)00011-3). (Cited on page [1](#).)
- [33] Neuweiler, H., Doose, S. & Sauer, M. A microscopic view of miniprotein folding: enhanced folding efficiency through formation of an intermediate. *Proc. Nat. Acad. Sci. USA* **102**, 16650–16655 (2005). doi [10.1073/pnas.0507351102](https://doi.org/10.1073/pnas.0507351102). (Cited on pages [1](#) and [29](#).)
- [34] Schaeffer, R. D., Fersht, A. & Daggett, V. Combining experiment and simulation in protein folding: closing the gap for small model systems. *Curr. Opin. Struct. Biol.* **18**, 4–9 (2008). doi [10.1016/j.sbi.2007.11.007](https://doi.org/10.1016/j.sbi.2007.11.007). (Cited on page [2](#).)
- [35] Noé, F., Schütte, C., Vanden-Eijnden, E. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Nat. Acad. Sci. USA* **106**, 19011–19016 (2009). doi [10.1073/pnas.0905466106](https://doi.org/10.1073/pnas.0905466106). (Cited on pages [2](#), [3](#), [4](#), [30](#), [34](#), [41](#), [42](#), [49](#), [51](#), [52](#), [114](#), [147](#), [148](#), and [150](#).)
- [36] van Gunsteren, W. F., Dolenc, J. & Mark, A. E. Molecular simulation as an aid to experimentalists. *Curr. Opin. Struct. Biol.* **18**, 149–153 (2008). doi [10.1016/j.sbi.2007.12.007](https://doi.org/10.1016/j.sbi.2007.12.007). (Cited on page [2](#).)
- [37] Krivov, S. V. & Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Nat. Acad. Sci. USA* **101**, 14766–14770 (2004). doi [10.1073/pnas.0406234101](https://doi.org/10.1073/pnas.0406234101). (Cited on pages [2](#) and [3](#).)
- [38] Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin.*

- Struct. Biol.* **18**, 154–162 (2008). doi [10.1016/j.sbi.2008.01.008](https://doi.org/10.1016/j.sbi.2008.01.008). (Cited on pages [2](#), [4](#), and [61](#).)
- [39] Muff, S. & Caflisch, A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *PROTEINS: Struct. Func. Bioinfo.* **70**, 1185–1195 (2008). doi [10.1002/prot.21565](https://doi.org/10.1002/prot.21565). (Cited on pages [2](#) and [3](#).)
- [40] Wales, D. J. Energy landscapes. Cambridge University Press, Cambridge (2003). (Cited on pages [2](#), [3](#), and [150](#).)
- [41] Karpen, M. E., Tobias, D. J. & Brooks, C. L. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry* **32**, 412–420 (1993). doi [10.1021/bi00053a005](https://doi.org/10.1021/bi00053a005). (Cited on page [2](#).)
- [42] Hubner, I. A., Deeds, E. J. & Shakhnovich, E. I. Understanding ensemble protein folding at atomic detail. *Proc. Nat. Acad. Sci. USA* **103**, 17747–17752 (2006). doi [10.1073/pnas.0605580103](https://doi.org/10.1073/pnas.0605580103). (Cited on page [2](#).)
- [43] Weber, M. Improved Perron Cluster Analysis. *ZIB Report* **04** (2003). url <http://www.zib.de/Publications/Reports/ZR-03-04.pdf>. (Cited on pages [2](#), [18](#), [106](#), [107](#), and [150](#).)
- [44] Buchete, N.-V. & Hummer, G. Coarse Master Equations for Peptide Folding Dynamics. *J. Phys. Chem. B* **112**, 6057–6069 (2008). doi [10.1021/jp0761665](https://doi.org/10.1021/jp0761665). (Cited on pages [2](#), [19](#), [30](#), [33](#), [34](#), [99](#), and [149](#).)
- [45] Rao, F. & Caflisch, A. The Protein Folding Network. *J. Mol. Biol.* **342**, 299–306 (2004). doi [10.1016/j.jmb.2004.06.063](https://doi.org/10.1016/j.jmb.2004.06.063). (Cited on page [2](#).)
- [46] Pan, A. & Roux, B. Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.* **129**, 064107 (2008). doi [10.1063/1.2959573](https://doi.org/10.1063/1.2959573). (Cited on pages [2](#), [3](#), and [148](#).)
- [47] Schultheis, V., Hirschberger, T., Carstens, H. & Tavan, P. Extracting Markov Models of Peptide Conformational Dynamics from Simulation Data. *J. Chem. Theo. Comp.* **1**, 515–526 (2005). doi [10.1021/ct050020x](https://doi.org/10.1021/ct050020x). (Cited on pages [2](#) and [3](#).)
- [48] de Groot, B. L., Daura, X., Mark, A. E. & Grubmüller, H. Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.* **309**, 299–313 (2001). doi [10.1006/jmbi.2001.4655](https://doi.org/10.1006/jmbi.2001.4655). (Cited on page [2](#).)

- [49] Sarich, M., Noé, F. & Schütte, C. On the approximation quality of Markov state models. *Multiscale Model. Sim.* **8**, 1154 (2010). doi [10.1137/090764049](https://doi.org/10.1137/090764049). (Cited on pages [2](#), [25](#), [27](#), [32](#), [33](#), and [99](#).)
- [50] Noé, F., Oswald, M., Reinelt, G., Fischer, S. & Smith, J. C. Computing Best Transition Pathways in High-Dimensional Dynamical Systems: Application to the Transitions in Octaalanine. *Multiscale Model. Sim.* **5**, 393–419 (2006). doi [10.1137/050641922](https://doi.org/10.1137/050641922). (Cited on pages [3](#) and [42](#).)
- [51] Noé, F., Horenko, I., Schütte, C. & Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **126**, 155102 (2007). doi [10.1063/1.2714539](https://doi.org/10.1063/1.2714539). (Cited on pages [3](#), [4](#), [18](#), [26](#), [29](#), [30](#), [33](#), [34](#), [47](#), [98](#), and [112](#).)
- [52] Chodera, J. D., Hinrichs, N. S., Pande, V. S., Dill, K. A. & Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101 (2007). doi [10.1063/1.2714538](https://doi.org/10.1063/1.2714538). (Cited on pages [3](#), [4](#), [26](#), [27](#), [30](#), [33](#), [34](#), [36](#), [47](#), [48](#), [98](#), [112](#), and [114](#).)
- [53] Swope, W. C., Pitera, J. W. & Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *J. Phys. Chem. B* **108**, 6571–6581 (2004). doi [10.1021/jp037421y](https://doi.org/10.1021/jp037421y). (Cited on pages [3](#), [4](#), [12](#), [17](#), [26](#), [29](#), [30](#), [32](#), [33](#), [40](#), [89](#), [99](#), [112](#), and [114](#).)
- [54] Hinrichs, N. S., Snow, C. D. & Pande, V. S. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* **121**, 415 (2004). doi [10.1063/1.1738647](https://doi.org/10.1063/1.1738647). (Cited on pages [3](#), [60](#), [61](#), and [62](#).)
- [55] Schütte, C., Fischer, A., Huisinga, W. & Deufflhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comp. Phys.* **151**, 146–168 (1999). doi [10.1006/jcph.1999.6231](https://doi.org/10.1006/jcph.1999.6231). (Cited on pages [3](#), [13](#), [14](#), [16](#), [32](#), and [150](#).)
- [56] Weber, M. A Subspace Approach to Molecular Markov State Models via an Infinitesimal Generator (revised version). *ZIB Report* **09**, 1–81 (2010). url <http://www.zib.de/Publications/Reports/ZR-09-27.pdf>. (Cited on pages [3](#) and [32](#).)
- [57] Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. The folding mechanism of a beta-sheet: the WW domain. *J. Mol.*

- Biol.* **311**, 373–393 (2001). doi [10.1006/jmbi.2001.4873](https://doi.org/10.1006/jmbi.2001.4873). (Cited on pages [3](#), [29](#), and [77](#).)
- [58] Chodera, J. D. & Noé, F. Probability distributions of molecular observables computed from Markov models. II. Uncertainties in observables and their time-evolution. *J. Chem. Phys.* **133**, 105102 (2010). doi [10.1063/1.3463406](https://doi.org/10.1063/1.3463406). (Cited on pages [3](#), [4](#), [30](#), [36](#), [42](#), and [150](#).)
- [59] Voelz, V. A., Bowman, G. R., Beauchamp, K. A. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010). doi [10.1021/ja9090353](https://doi.org/10.1021/ja9090353). (Cited on pages [3](#) and [114](#).)
- [60] Hinrichs, N. S. & Pande, V. S. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **123**, 204909 (2005). doi [10.1063/1.2116947](https://doi.org/10.1063/1.2116947). (Cited on pages [4](#), [41](#), [42](#), [60](#), [61](#), [62](#), [63](#), [95](#), [149](#), and [150](#).)
- [61] Hinrichs, N. S. & Pande, V. S. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.* **126**, 244101 (2007). doi [10.1063/1.2740261](https://doi.org/10.1063/1.2740261). (Cited on pages [4](#), [42](#), [91](#), and [149](#).)
- [62] Bowman, G. R., Beauchamp, K. A., Boxer, G. & Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **131**, 124101 (2009). doi [10.1063/1.3216567](https://doi.org/10.1063/1.3216567). (Cited on pages [4](#), [34](#), [40](#), and [112](#).)
- [63] Metzner, P., Schütte, C. & Vanden-Eijnden, E. Transition path theory for Markov jump processes. *Multiscale Model. Sim.* **7**, 1192–1219 (2009). doi [10.1137/070699500](https://doi.org/10.1137/070699500). (Cited on pages [4](#), [51](#), [52](#), [54](#), [73](#), and [150](#).)
- [64] Noé, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **128**, 244103 (2008). doi [10.1063/1.2916718](https://doi.org/10.1063/1.2916718). (Cited on pages [4](#), [34](#), [36](#), [42](#), [61](#), [62](#), [83](#), [84](#), and [87](#).)
- [65] Chodera, J. D., Swope, W. C., Pitera, J. W. & Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Sim.* **5**, 1214–1226 (2006). doi [10.1137/06065146X](https://doi.org/10.1137/06065146X). (Cited on pages [4](#), [26](#), [78](#), [82](#), [84](#), [85](#), [86](#), [87](#), [98](#), and [147](#).)
- [66] Taketomi, H., Ueda, Y. & Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int J Pept Prot Res* **7**, 445–459 (1975). doi [10.1111/j.1399-3011.1975.tb02465.x](https://doi.org/10.1111/j.1399-3011.1975.tb02465.x). (Cited on page [8](#).)

- [67] Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72**, 2384 (1980). doi [10.1063/1.439486](https://doi.org/10.1063/1.439486). (Cited on pages [11](#), [12](#), [83](#), and [157](#).)
- [68] Ermak, D. L. & Yeh, Y. Equilibrium electrostatic effects on the behavior of polyions in solution: polyion-mobile ion interaction. *Chem. Phys. Lett.* **24**, 243–248 (1974). doi [10.1016/0009-2614\(74\)85442-4](https://doi.org/10.1016/0009-2614(74)85442-4). (Cited on page [12](#).)
- [69] Ermak, D. L. A computer simulation of charged particles in solution. I. Technique and equilibrium properties. *J. Chem. Phys.* **62**, 4189–4196 (1975). doi [10.1063/1.430300](https://doi.org/10.1063/1.430300). (Cited on page [12](#).)
- [70] Cooke, B. & Schmidler, S. C. Statistical prediction and molecular dynamics simulation. *Biophys. J.* **95**, 4497–4511 (2008). doi [10.1529/biophysj.108.131623](https://doi.org/10.1529/biophysj.108.131623). (Cited on page [12](#).)
- [71] Swope, W. C., Andersen, H. C., Berens, P. & Wilson, K. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76**, 637 (1982). doi [10.1063/1.442716](https://doi.org/10.1063/1.442716). (Cited on page [12](#).)
- [72] Schütte, C., Noé, F., Meerbach, E., Metzner, P. & Hartmann, C. Conformation dynamics. *ICIAM 07: 6th International Conference on Industrial and Applied Mathematics, Zürich, Switzerland, 16-20 July 2007: invited lectures* 297 (2009). (Cited on pages [13](#) and [32](#).)
- [73] Sriraman, S., Kevrekidis, I. G. & Hummer, G. Coarse Master Equation from Bayesian Analysis of Replica Molecular Dynamics Simulations. *J. Phys. Chem. B* **109**, 6479–6484 (2005). doi [10.1021/jp046448u](https://doi.org/10.1021/jp046448u). (Cited on pages [19](#), [33](#), and [149](#).)
- [74] van Kampen, N. G. *Stochastic Processes in Physics and Chemistry* (Elsevier, 2006). (Cited on pages [19](#) and [149](#).)
- [75] Röblitz, S. & Weber, M. A coarse graining method for the identification of transition rates between molecular conformations. *J. Chem. Phys.* **126**, 024103 (2007). doi [10.1063/1.2404953](https://doi.org/10.1063/1.2404953). (Cited on pages [26](#), [32](#), and [99](#).)
- [76] Chan, C. K., Hu, Y., Takahashi, S., Rousseau, D. L., Eaton, W. A. & Hofrichter, J. Submillisecond protein folding kinetics studied by ultrarapid mixing. *Proc. Nat. Acad. Sci. USA* **94**, 1779–1784 (1997). doi [10.1073/pnas.94.5.1779](https://doi.org/10.1073/pnas.94.5.1779). (Cited on page [29](#).)
- [77] Bieri, O., Wirz, J., Hellrung, B., Schutkowski, M., Drewello, M. & Kiefhaber, T. The speed limit for protein folding measured by triplet-triplet energy transfer. *Proc. Nat. Acad. Sci. USA* **96**, 9597–9601 (1999). doi [10.1073/pnas.96.17.9597](https://doi.org/10.1073/pnas.96.17.9597). (Cited on page [29](#).)

- [78] Djurdjevac, N., Sarich, M. & Schütte, C. Estimating the Eigenvalue Error of Markov State Models. *Multiscale Model. Sim.* **10**, 61–81 (2012). doi [10.1137/100798910](https://doi.org/10.1137/100798910). (Cited on pages 30 and 150.)
- [79] Swope, W. C. *et al.* Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 2. Example Applications to Alanine Dipeptide and a beta-Hairpin Peptide. *J. Phys. Chem. B* **108**, 6582–6594 (2004). doi [10.1021/jp037422q](https://doi.org/10.1021/jp037422q). (Cited on pages 30, 33, 34, 40, and 89.)
- [80] Crommelin, D. & Vanden-Eijnden, E. Data-Based Inference of Generators for Markov Jump Processes Using Convex Optimization. *Multiscale Model. Sim.* **7**, 1751–1778 (2009). doi [10.1137/080735977](https://doi.org/10.1137/080735977). (Cited on page 33.)
- [81] Metzner, P., Horenko, I. & Schütte, C. Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time. *Phys. Rev. E* **76**, 066702 (2007). doi [10.1103/PhysRevE.76.066702](https://doi.org/10.1103/PhysRevE.76.066702). (Cited on page 33.)
- [82] Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential dynamics of proteins. *PROTEINS* **17**, 412–425 (1993). doi [10.1002/prot.340170408](https://doi.org/10.1002/prot.340170408). (Cited on page 33.)
- [83] Chodera, J. D., Swope, W. C., Pitera, J. W., Seok, C. & Dill, K. A. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *J. Chem. Theo. Comp.* **3**, 26–41 (2007). doi [10.1021/ct0502864](https://doi.org/10.1021/ct0502864). (Cited on pages 34 and 86.)
- [84] Keller, B. G., Daura, X. & van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **132**, 074110 (2010). doi [10.1063/1.3301140](https://doi.org/10.1063/1.3301140). (Cited on page 34.)
- [85] Yao, Y., Sun, J., Huang, X., Bowman, G. R., Singh, G., Lesnick, M., Guibas, L. J., Pande, V. S. & Carlsson, G. Topological methods for exploring low-density states in biomolecular folding pathways. *J. Chem. Phys.* **130**, 144115 (2009). doi [10.1063/1.3103496](https://doi.org/10.1063/1.3103496). (Cited on page 34.)
- [86] Meerbach, E., Schütte, C., Horenko, I. & Schmidt, B. Metastable Conformational Structure and Dynamics: Peptides between Gas Phase and Aqueous Solution. In *Analysis and Control of Ultrafast Photoinduced Reactions. Series in Chemical Physics*, 796–806 (Springer, Berlin, 2007). (Cited on page 34.)
- [87] Anderson, T. W. & Goodman, L. A. Statistical Inference about Markov Chains. *Ann. Math. Stat.* **28**, 89–110 (1957). doi [10.1214/aoms/1177707039](https://doi.org/10.1214/aoms/1177707039). (Cited on pages 37 and 40.)

- [88] Jaynes, E. T. *Probability Theory: The Logic of Science* (Cambridge University Press, 2003). (Cited on page 38.)
- [89] Noé, F., Oswald, M. & Reinelt, G. Optimizing in graphs with expensive computation of edge weights. *Operations Research Proceedings 2007: Selected Papers of the Annual International Conference of the German Operations Research Society (Gor)* 435 (2008). (Cited on page 42.)
- [90] Metzner, P., Noé, F. & Schütte, C. Estimating the sampling error: distribution of transition matrices and functions of transition matrices for given trajectory data. *Physical review E, Statistical, nonlinear, and soft matter physics* **80**, 021106 (2009). doi [10.1103/PhysRevE.80.021106](https://doi.org/10.1103/PhysRevE.80.021106). (Cited on page 42.)
- [91] Lange, O. F. & Grubmüller, H. Collective Langevin dynamics of conformational motions in proteins. *J. Chem. Phys.* **124**, 214903 (2006). doi [10.1063/1.2199530](https://doi.org/10.1063/1.2199530). (Cited on pages 47 and 150.)
- [92] E, W., Ren, W. & Vanden-Eijnden, E. Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* **413**, 242–247 (2005). doi [10.1016/j.cplett.2005.07.084](https://doi.org/10.1016/j.cplett.2005.07.084). (Cited on pages 49 and 50.)
- [93] Dellago, C., Bolhuis, P. G. & Geissler, P. L. Transition path sampling. *Adv. Chem. Phys.* **123**, 1 (2002). doi [10.1002/0471231509.ch1](https://doi.org/10.1002/0471231509.ch1). (Cited on page 49.)
- [94] Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **53**, 291 (2002). doi [10.1146/annurev.physchem.53.082301.113146](https://doi.org/10.1146/annurev.physchem.53.082301.113146). (Cited on page 49.)
- [95] Maragliano, L., Fischer, A., Vanden-Eijnden, E. & Ciccotti, G. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **125**, 24106 (2006). doi [10.1063/1.2212942](https://doi.org/10.1063/1.2212942). (Cited on page 49.)
- [96] Best, R. B. & Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Nat. Acad. Sci. USA* **102**, 6732–6737 (2005). doi [10.1073/pnas.0408098102](https://doi.org/10.1073/pnas.0408098102). (Cited on pages 49 and 50.)
- [97] Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. I. On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334 (1998). doi [10.1063/1.475393](https://doi.org/10.1063/1.475393). (Cited on page 49.)
- [98] Hummer, G. From transition paths to transition states and rate coefficients. *J. Chem. Phys.* **120**, 516–523 (2004). doi [10.1063/1.1630572](https://doi.org/10.1063/1.1630572). (Cited on page 49.)

- [99] Ma, A. & Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* **109**, 6769–6779 (2005). doi [10.1021/jp045546c](https://doi.org/10.1021/jp045546c). (Cited on pages 49 and 50.)
- [100] Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68–79 (1998). doi [10.1016/S0959-440X\(98\)80012-2](https://doi.org/10.1016/S0959-440X(98)80012-2). (Cited on page 51.)
- [101] Vanden-Eijnden, E. *Transition Path Theory*, vol. 703 (Springer Verlag, 2006). doi [10.1007/3-540-35273-2](https://doi.org/10.1007/3-540-35273-2). (Cited on pages 51 and 52.)
- [102] Wales, D. J. Energy landscapes: calculating pathways and rates. *International Reviews in Physical Chemistry* **25**, 237–282 (2006). doi [10.1080/01442350600676921](https://doi.org/10.1080/01442350600676921). (Cited on page 52.)
- [103] Norris, J. R. *Markov chains* (Cambridge University Press, 1998). (Cited on page 54.)
- [104] Lenz, P., Zagrovic, B., Shapiro, J. & Pande, V. S. Folding probabilities: A novel approach to folding transitions and the two-dimensional Ising-model. *J. Chem. Phys.* **120**, 6769 (2004). doi [10.1063/1.1667470](https://doi.org/10.1063/1.1667470). (Cited on page 55.)
- [105] Berezhkovskii, A. M. & Szabo, A. Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *J. Chem. Phys.* **121**, 9186 (2004). doi [10.1063/1.1802674](https://doi.org/10.1063/1.1802674). (Cited on page 55.)
- [106] Golub, G. & van Loan, C. *Matrix computation, third Edition* (John Hopkins University Press, 1996). (Cited on pages 58, 59, and 72.)
- [107] Wales, D. J. Calculating rate constants and committor probabilities for transition networks by graph transformation. *J. Chem. Phys.* **130**, 204111 (2009). doi [10.1063/1.3133782](https://doi.org/10.1063/1.3133782). (Cited on page 59.)
- [108] Schütte, C., Noé, F., Lu, J., Sarich, M. & Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **134**, 204105 (2011). doi [10.1063/1.3590108](https://doi.org/10.1063/1.3590108). (Cited on pages 60, 99, and 150.)
- [109] Rogal, J. & Bolhuis, P. G. Multiple state transition path sampling. *J. Chem. Phys.* **129**, 224107 (2008). doi [10.1063/1.3029696](https://doi.org/10.1063/1.3029696). (Cited on page 60.)
- [110] Geyer, C. J. & Thompson, E. A. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*

- 90, 909–920 (1995). doi [10.1080/01621459.1995.10476590](https://doi.org/10.1080/01621459.1995.10476590). (Cited on page 77.)
- [111] Hukushima, K. & Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan* **65**, 1604–1608 (1996). doi [10.1143/JPSJ.65.1604](https://doi.org/10.1143/JPSJ.65.1604). (Cited on page 77.)
- [112] Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **281**, 140–150 (1997). doi [10.1016/S0009-2614\(97\)01198-6](https://doi.org/10.1016/S0009-2614(97)01198-6). (Cited on page 77.)
- [113] Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999). doi [10.1016/S0009-2614\(99\)01123-9](https://doi.org/10.1016/S0009-2614(99)01123-9). (Cited on pages 77, 82, 83, and 148.)
- [114] Buchete, N.-V. & Hummer, G. Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E* **77**, 4 (2008). doi [10.1103/PhysRevE.77.030902](https://doi.org/10.1103/PhysRevE.77.030902). (Cited on pages 77, 78, 79, 84, 86, 87, and 148.)
- [115] Gallicchio, E., Andrec, M., Felts, A. K. & Levy, R. M. Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths. *J. Phys. Chem. B* **109**, 6722–6731 (2005). doi [10.1021/jp045294f](https://doi.org/10.1021/jp045294f). (Cited on pages 78 and 95.)
- [116] Ferrenberg, A. & Swendsen, R. H. Optimized Monte Carlo data analysis. *Phys. Rev. Lett* **63**, 1195–1198 (1989). doi [10.1103/PhysRevLett.63.1195](https://doi.org/10.1103/PhysRevLett.63.1195). (Cited on page 78.)
- [117] Bartels, C. & Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comp. Chem.* **18**, 1450–1462 (1997). doi [10.1002/\(SICI\)1096-987X\(199709\)18:12<1450::AID-JCC3>3.0.CO;2-I](https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1450::AID-JCC3>3.0.CO;2-I). (Cited on page 78.)
- [118] Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. The weighted histogram analysis method for free-energy calculations on biomolecules. I: The method. *J. Comp. Chem.* **13**, 1011–1021 (1992). doi [10.1002/jcc.540130812](https://doi.org/10.1002/jcc.540130812). (Cited on pages 78, 86, and 155.)
- [119] Bartels, C. Analyzing biased Monte Carlo and molecular dynamics simulations. *Chem. Phys. Lett.* **331**, 446–454 (2000). doi [10.1016/S0009-2614\(00\)01215-X](https://doi.org/10.1016/S0009-2614(00)01215-X). (Cited on page 78.)
- [120] Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**,

- 124105 (2008). doi [10.1063/1.2978177](https://doi.org/10.1063/1.2978177). (Cited on pages [78](#), [80](#), [81](#), [82](#), [155](#), and [156](#).)
- [121] Muff, S. & Caflisch, A. ETNA: equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations. *J. Phys. Chem. B* **113**, 3218–3226 (2009). doi [10.1021/jp807261h](https://doi.org/10.1021/jp807261h). (Cited on page [78](#).)
- [122] Mitsutake, A., Sugita, Y. & Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **60**, 96–123 (2001). doi [10.1002/1097-0282\(2001\)60:2<96::AID-BIP1007>3.0.CO;2-F](https://doi.org/10.1002/1097-0282(2001)60:2<96::AID-BIP1007>3.0.CO;2-F). (Cited on page [78](#).)
- [123] Earl, D. J. & Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **7**, 3910 (2005). doi [10.1039/b509983h](https://doi.org/10.1039/b509983h). (Cited on page [82](#).)
- [124] Verlet, L. Computer "Experiments" on Classical Fluids. II. Equilibrium Correlation Functions. *Phys. Rev.* **165**, 201–214 (1968). doi [10.1103/PhysRev.165.201](https://doi.org/10.1103/PhysRev.165.201). (Cited on page [85](#).)
- [125] Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **159**, 98–103 (1967). doi [10.1103/PhysRev.159.98](https://doi.org/10.1103/PhysRev.159.98). (Cited on page [85](#).)
- [126] Ryckaert, J., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* **23**, 327–341 (1977). doi [10.1016/0021-9991\(77\)90098-5](https://doi.org/10.1016/0021-9991(77)90098-5). (Cited on page [85](#).)
- [127] Bowman, G. R., Ensign, D. L. & Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theo. Comp.* **6**, 787–794 (2010). doi [10.1021/ct900620b](https://doi.org/10.1021/ct900620b). (Cited on pages [92](#) and [149](#).)
- [128] Greenleaf, W. J., Woodside, M. T. & Block, S. M. High-resolution, single-molecule measurements of biomolecular motion. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 171–190 (2007). doi [10.1146/annurev.biophys.36.101106.101451](https://doi.org/10.1146/annurev.biophys.36.101106.101451). (Cited on page [97](#).)
- [129] Lapidus, L. J., Eaton, W. A. & Hofrichter, J. Measuring the rate of intramolecular contact formation in polypeptides. *Proc. Nat. Acad. Sci. USA* **97**, 7220–7225. (2000). doi [10.1073/pnas.97.13.7220](https://doi.org/10.1073/pnas.97.13.7220). (Cited on page [97](#).)

- [130] Zhou, H.-X. Rate theories for biologists. *Quart. Rev. Biophys.* **43**, 219–293 (2010). doi [10.1017/S0033583510000120](https://doi.org/10.1017/S0033583510000120). (Cited on pages [97](#), [98](#), and [111](#).)
- [131] Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **3**, 107 (1935). doi [10.1063/1.1749604](https://doi.org/10.1063/1.1749604). (Cited on page [98](#).)
- [132] Chandler, D. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.* **68**, 2959–2970 (1978). doi [10.1063/1.436049](https://doi.org/10.1063/1.436049). (Cited on page [98](#).)
- [133] Chodera, J. D., Elms, P. J., Swope, W. C., Prinz, J.-H., Marqusee, S., Bustamante, C., Noé, F. & Pande, V. S. A robust approach to estimating rates from time-correlation functions. *Arxiv preprint cond-mat* (2011). url <http://arxiv.org/abs/1108.2304>. (Cited on pages [98](#), [112](#), [113](#), and [169](#).)
- [134] Elson, E. L. & Magde, D. Fluorescence Correlation Spectroscopy. I. Conceptual Basis and Theory. *Biopolymers* **13**, 1–27 (1974). doi [10.1002/bip.1974.360130102](https://doi.org/10.1002/bip.1974.360130102). (Cited on page [98](#).)
- [135] Skinner, J. L. & Wolynes, P. G. Relaxation processes and chemical kinetics. *J. Chem. Phys.* **69**, 2143 (1978). doi [10.1063/1.436814](https://doi.org/10.1063/1.436814). (Cited on pages [98](#), [111](#), and [113](#).)
- [136] McKinney, S. & Joo, C. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophys. J.* (2006). doi [10.1529/biophysj.106.082487](https://doi.org/10.1529/biophysj.106.082487). (Cited on page [99](#).)
- [137] Gopich, I. V., Nettels, D., Schuler, B. & Szabo, A. Protein dynamics from single-molecule fluorescence intensity correlation functions. *J. Chem. Phys.* **131**, 095102 (2009). doi [10.1063/1.3212597](https://doi.org/10.1063/1.3212597). (Cited on page [99](#).)
- [138] Chodera, J. D., Elms, P., Noé, F., Keller, B. G., Kaiser, C. M., Ewall-Wice, A., Marqusee, S., Bustamante, C. & Hinrichs, N. S. Bayesian hidden Markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty. *Arxiv preprint cond-mat* (2011). url <http://arxiv.org/abs/1108.1430>. (Cited on page [99](#).)
- [139] Noé, F., Doose, S., Daidone, I., Löllmann, M., Sauer, M., Chodera, J. D. & Smith, J. C. Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc. Nat. Acad. Sci. USA* **108**, 4822–4827 (2011). doi [10.1073/pnas.1004646108](https://doi.org/10.1073/pnas.1004646108). (Cited on page [104](#).)
- [140] Ritz, W. *Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik*, vol. 1909 (Journal für die

- reine und angewandte Mathematik (Crelle's Journal), 1909). doi [10.1515/crll.1909.135.1](https://doi.org/10.1515/crll.1909.135.1). (Cited on page 109.)
- [141] Peters, B. & Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **125**, 054108 (2006). doi [10.1063/1.2234477](https://doi.org/10.1063/1.2234477). (Cited on page 109.)
- [142] Zwanzig, R. & Ailawadi, N. Statistical Error Due to Finite Time Averaging in Computer Experiments. *Phys. Rev.* **182**, 280–283 (1969). doi [10.1103/PhysRev.182.280](https://doi.org/10.1103/PhysRev.182.280). (Cited on page 111.)
- [143] Noé, F. & Nüske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Arxiv preprint math-ph* (2012). url <http://arxiv.org/abs/1211.7103>. (Cited on page 111.)
- [144] Onsager, L. Initial Recombination of Ions. *Phys. Rev.* **54**, 554–557 (1938). doi [10.1103/PhysRev.54.554](https://doi.org/10.1103/PhysRev.54.554). (Cited on page 112.)
- [145] Demmel, J. Generalized Hermitian Eigenproblems . In Bai, Z., Demmel, J., Dongarra, J., Ruhe, A. & van der Vorst, H. (eds.) *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide* (SIAM, Philadelphia, 2000). (Cited on page 117.)
- [146] de Leeuw, J. Derivatives of Generalized Eigen Systems with Applications (2007). url <http://statistics.ucla.edu/preprints/uclastat-preprint-2007:18>. (Cited on page 119.)
- [147] Elms, P. J., Chodera, J. D., Bustamante, C. J. & Marqusee, S. Limitations of constant-force-feedback experiments. *Biophys. J.* **103**, 1490–1499 (2012). doi [10.1016/j.bpj.2012.06.051](https://doi.org/10.1016/j.bpj.2012.06.051). (Cited on pages 126, 127, 129, 169, and 170.)
- [148] Wolfram Research. *Mathematica* (Wolfram Research, Champaign, Illinois, 2010), version 8.0 edn. (Cited on page 127.)
- [149] Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nuc. Acid. Res.* **36**, W70–W74 (2008). doi [10.1093/nar/gkn188](https://doi.org/10.1093/nar/gkn188). (Cited on page 130.)
- [150] Wang, M. D., Yin, H., Landick, R., Gelles, J. & Block, S. M. Stretching DNA with optical tweezers. *Biophys. J.* **72**, 1335–1346 (1997). doi [10.1016/S0006-3495\(97\)78780-0](https://doi.org/10.1016/S0006-3495(97)78780-0). (Cited on page 133.)
- [151] Manosas, M. & Ritort, F. Thermodynamic and Kinetic Aspects of RNA Pulling Experiments. *Biophys. J.* **88**, 3224–3242 (2005). doi [10.1529/biophysj.104.045344](https://doi.org/10.1529/biophysj.104.045344). (Cited on page 133.)
- [152] Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Nat. Acad. Sci. USA* **99**, 12562 (2002).

- url <http://www.pnas.org/cgi/content/abstract/99/20/12562>. (Cited on page 148.)
- [153] Grubmüller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E* **52**, 2893–2906 (1995). doi [10.1103/PhysRevE.52.2893](https://doi.org/10.1103/PhysRevE.52.2893). (Cited on page 148.)
- [154] Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.* **23**, 187–199 (1977). doi [10.1016/0021-9991\(77\)90121-8](https://doi.org/10.1016/0021-9991(77)90121-8). (Cited on page 148.)
- [155] Schlitter, J., Engels, M. & Krüger, P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *Journal of molecular graphics* **12**, 84–89 (1994). doi [10.1016/0263-7855\(94\)80072-3](https://doi.org/10.1016/0263-7855(94)80072-3). (Cited on page 148.)
- [156] Wales, D. J. Structure, Dynamics, and Thermodynamics of Clusters: Tales from Topographic Potential Surfaces. *Science* **271**, 925–929 (1996). doi [10.1126/science.271.5251.925](https://doi.org/10.1126/science.271.5251.925). (Cited on page 150.)
- [157] Hegger, R. & Stock, G. Multidimensional Langevin modeling of biomolecular dynamics. *J. Chem. Phys.* **130**, 034106 (2009). doi [10.1063/1.3058436](https://doi.org/10.1063/1.3058436). (Cited on page 150.)
- [158] Micheletti, C., Bussi, G. & Laio, A. Optimal Langevin modeling of out-of-equilibrium molecular dynamics simulations. *J. Chem. Phys.* **129**, 074105 (2008). doi [10.1063/1.2969761](https://doi.org/10.1063/1.2969761). (Cited on page 150.)
- [159] Khinchin, A. I. *Mathematical Foundations of Statistical Mechanics* (Dover Publications, 1949), first edn. (Cited on page 157.)
- [160] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (1953). doi [10.1063/1.1699114](https://doi.org/10.1063/1.1699114). (Cited on page 168.)
- [161] Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I. & Bustamante, C. Reversible unfolding of single RNA molecules by mechanical force. *Science* **292**, 733–737 (2001). doi [10.1126/science.1058498](https://doi.org/10.1126/science.1058498). (Cited on page 169.)
- [162] Wen, J.-D., Manosas, M., Li, P. T. X., Smith, S. B., Bustamante, C., Ritort, F. & Tinoco Jr, I. Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results. *Biophys. J.* **92**, 2996–3009 (2007). doi [10.1529/biophysj.106.094052](https://doi.org/10.1529/biophysj.106.094052). (Cited on page 169.)

- [163] Smith, S. B., Cui, Y. & Bustamante, C. Optical-trap force transducer that operates by direct measurement of light momentum. *Meth. Enzymol.* **361**, 134–162 (2003). doi [10.1016/S0076-6879\(03\)61009-8](https://doi.org/10.1016/S0076-6879(03)61009-8). (Cited on page [169](#).)
- [164] Bustamante, C. & Smith, S. B. Light-force sensor and method for measuring axial optical-trap forces from changes in light momentum along an optical axis. *United States Patent No. 7133132* (2006). (Cited on page [169](#).)
- [165] Bringhurst, R. *The Elements of Typographic Style*. Version 3.2 (Hartley & Marks Publishers, Point Roberts, WA, USA, 2008). (Cited on page [189](#).)



## COLOPHON

This thesis was typeset with  $\text{\LaTeX} 2_{\epsilon}$  using Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URW Palladio L* and *FPL* were used).

The typographic style was inspired by Robert Bringhurst as presented in *The Elements of Typographic Style* [165]. It is available for  $\text{\LaTeX}$  via CTAN as "**classicthesis**".