TECHNISCHE UNIVERSITÄT DARMSTADT

BAYESIAN INFERENCE AND LEARNING IN SWITCHING BIOLOGICAL SYSTEMS

Vom Fachbereich Elektrotechnik und Informationstechnik der
TECHNISCHEN UNIVERSITÄT DARMSTADT

zur Erlangung des akademischen Grades eines
Doctor rerum naturalium (Dr. rer. nat.)
genehmigte Dissertation

von

LUKAS KÖHS, M.SC.

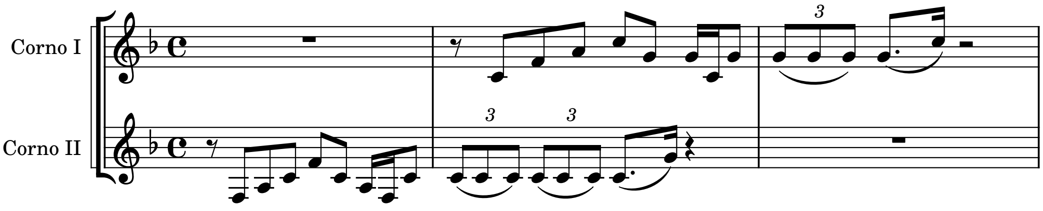
Referent:	Prof. Dr. techn. Heinz Köppl
Korreferent:	Prof. Dr. Manfred Opper

Tag der Einreichung:	17. Oktober 2022
Tag der mündlichen Prüfung:	15. Dezember 2022

D17
Darmstadt 2022

Die Arbeit von Lukas Köhs wurde vom European Research Council (ERC) im Rahmen des Projektes 773196 „CONSYN – Contextualizing biomolecular circuits models for synthetic biology“ gefördert.

Köhs, Lukas: *Bayesian Inference and Learning in Switching Biological Systems*
Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2023
URN: urn:nbn:de:tuda-tuprints-230220
Tag der mündlichen Prüfung: 15. Dezember 2022



— Johann Sebastian Bach

„Jede Idee entsteht im Dialog.“

— Nikolaus Harnoncourt

KURZFASSUNG

Diese Dissertation befasst sich mit der stochastischen Modellierung schaltender biologischer Systeme und der Entwicklung entsprechender Inferenzalgorithmen. Ausgehend von der großen Vielfalt der Mess- und Simulationsmethoden, die zur Analyse solcher Systeme zur Anwendung kommen, werden sowohl zeitkontinuierliche als auch zeitdiskrete Modellierungsansätze verfolgt. Weiterhin werden einerseits hybride, kontinuierlich-diskrete, andererseits rein diskrete latente Zustandsräume betrachtet. Für die Zeitdynamik der betrachteten Systeme sowie für ihre Parameter werden Bayes'sche Inferenzmethoden entwickelt: Ausgehend von der exakten Modellformulierung werden jeweils Approximationen abgeleitet, die zu rechnerisch handhabbaren Algorithmen führen. Diese Approximationen basieren entweder auf Sampling- oder auf Variationsprinzipien. Die so formulierten Algorithmen werden dann sowohl unter der jeweiligen Modellannahme getestet wie auch nachfolgend auf bekannte Benchmarkprobleme und experimentalbiologische Daten angewandt. Die Arbeit gliedert sich dabei in drei wissenschaftliche Beiträge:

Erstens wird eine *Markov chain Monte Carlo*-Methode für zeitkontinuierliche Prozesse mit hybridem Zustandsraum vorstellt. Diese Hybridprozesse werden als Markov-schaltende stochastische Differentialgleichungen formuliert, für die eine exakte Evolutionsgleichung hergeleitet werden kann. Um daraus eine rechnerisch handhabbare Inferenzmethode zu entwickeln, wird ein *Gibbs sampling*-Ansatz verwendet, der es erlaubt, sowohl die Zustandsdynamik wie auch die Systemparameter abzuschätzen. Dieser Ansatz wird dann unter der Modellannahme validiert und auf biologische Echtdaten eines genetischen Schaltexperimentes angewendet.

Zweitens wird ein Variationsansatz für das gleiche Problem hergeleitet, um die für die Inferenz nötigen Rechenlaufzeiten zu verkürzen. Dazu wird zunächst die Kullback-Leibler-Divergenz zwischen zwei echten schaltenden stochastischen Differentialgleichungen hergeleitet. Das Variationsmaß wird dann als Mischverteilung von Gaußprozessen formuliert, die eine schaltende stochastische Differentialgleichung approximiert, und es wird gezeigt, in welchem Regime diese Näherung Gültigkeit hat. Schließlich wird der Variationsansatz auf den gleichen synthetischen Daten wie die Samplingmethode getestet und auf Modellsysteme aus der rechnergestützten Strukturbiologie angewandt.

Drittens wird ein nichtparametrischer Inferenzalgorithmus für den Konformationswechsel von Molekülen vorgestellt. Hier wird ein rein diskretwertiger latenter Zustandsraum zugrunde gelegt, wobei jeder latente Zustand einer Molekülstruktur entspricht. Unter der erneuten Verwendung von Variationsprinzipien wird eine Approximation vorgestellt, um die Anzahl latenter Konformationen aus Daten zu schätzen. Diese Methode verallgemeinert den Ansatz des *Markov state modeling*, der seit geraumer Zeit in der rechnergestützten Strukturbiologie etabliert ist. Dazu wird ein Observationsmodell eingeführt, das für strukturelle Moleküldaten besonders gut geeignet ist. Um den Inferenzalgorithmus praktisch berechenbar zu machen, wird an dieser Stelle eine zweite Approximation vorgenommen. Schließlich wird auch dieser Ansatz sowohl unter der Modellannahme validiert als auch für bekannte Probleme aus der Strukturbiologie verwendet.

ABSTRACT

This thesis is concerned with the stochastic modeling of and inference for switching biological systems. Motivated by the great variety of data obtainable from such systems by wet-lab experiments or computer simulations, continuous-time as well as discrete-time frameworks are devised. Similarly, different latent state-space configurations - both hybrid continuous-discrete and purely discrete state spaces - are considered. These models enable Bayesian inferences about the temporal system dynamics as well as the respective parameters. Starting with the exact model formulations, principled approximations are derived using sampling and variational techniques, enabling computationally tractable algorithms. The resulting frameworks are evaluated under the modeling assumption and subsequently applied to common benchmark problems and real-world biological data. These developments are divided into three scientific contributions:

First, a Markov chain Monte Carlo method for continuous-time and continuous-discrete state-space hybrid processes is derived. These hybrid processes are formulated as Markov-switching stochastic differential equations, for which the exact evolution equation is also presented. A Gibbs sampling scheme is then derived which enables tractable inference both for the system dynamics and the system parameters. This approach is validated under the modeling assumption as well as applied to data from a wet-lab gene-switching experiment.

Second, a variational approach to the same problem is taken to speed up the inference procedure. To this end, a mixture of Gaussian processes serves as the variational measure. The method is derived starting from the Kullback-Leibler divergence between two true switching stochastic differential equations, and it is shown in which regime the Gaussian mixture approximation is valid. It is then benchmarked on the same ground-truth data as the Gibbs sampler and applied to model systems from computational structural biology.

Third, a nonparametric inference framework is laid out for conformational molecule switching. Here, a purely discrete latent state space is assumed, where each latent state corresponds to one molecular structure. Utilizing variational techniques again, a method is presented to identify the number of conformations present in the data. This method generalizes the framework of Markov state models, which is well-established in the field of computational structural biology. An observation likelihood model tailored to structural molecule data is introduced, along with an suitable approximation enabling tractable inference. This framework, too, is first evaluated on data generated under the model assumption and then applied to common problems in the field.

PUBLICATIONS

The following manuscripts were published during the period of the doctoral candidacy:

[1] L. Köhs, B. Alt, and H. Koeppl, “Markov chain Monte Carlo for continuous-time switching dynamical systems”, *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, pp. 11 430–11 454, 2022

[2] L. Köhs, B. Alt, and H. Koeppl, “Variational inference for continuous-time switching dynamical systems”, *Advances in Neural Information Processing Systems*, vol. 34, 2021

[3] L. Köhs, K. Kukovetz, O. Rauh, and H. Koeppl, “Nonparametric Bayesian inference for meta-stable conformational dynamics”, *Physical Biology*, vol. 19, no. 5, p. 056 006, 2022

ACKNOWLEDGMENTS

I am sincerely grateful to my advisor Professor Heinz Koeppl for providing me with the opportunity to work in the intriguing research field that is emerging at the intersection between mathematics, biology and engineering. It is at this intersection that his Self-Organizing Systems lab is placed, and I want to express my deep appreciation for his tireless efforts to make it a unique place where theory and experiment are brought together in a joint research effort. In this stimulating environment, he let me pursue my research interests freely and provided me with continuous support throughout my doctoral studies.

I am also deeply indebted to my colleague, collaborator and mentor of sorts Bastian Alt. His great enthusiasm, oversight and communicative skills have helped me to overcome the unavoidable hardships during a PhD endeavour and stay on course. His support and that of the other colleagues of my "generation", Dominik Linzner, Christian Wildner and Tim Prangemeier as well as Adrian Šošić, have been invaluable. Much (meshin) learn, such pleasure, very fun - BCSlers doin' me a heckin amaze! WOW!

I further thank the remaining current and former members of the Koeppl lab that I had the pleasure of working with: Ahmed Elshamanhory, Alina Kuzembayeva, Derya Altıntan, Felix Reinhardt, François Lehr, Gamze Dogali, Yannick Eich, Christian Fabian, Sascha Hauck, Arsen Korpetayev, Gizem Ekinci, Hongfei Liu, Jascha Diemer, Kai Cui, Kilian Heck, Klaus-Dieter Voss, Leo Bronstein, Matthias Schultheis, Anam Tahir, Wasiur KhudaBukhsh, Maik Molderings, Maleen Hanst, Mark Sinzger, Markus Röder, Megan Bailey, Melanie Mikosch-Wersching, Mengguang Li, Michael Schmidt, Miloš Lješković, Nicolai Engelmann, Nikita Kruk, Özdemir Cetin, Philipp Fröhlich, Rogier Schoeman, Sikun Yang, Sofia Startceva, Stanislav Stepaniuk and Yujie Zhong.

I also want to gratefully acknowledge that the (too often neglected) issues of organization and infrastructure are taken seriously in the Koeppl lab. A vast amount of work hours and stress is saved for everyone in the group by system administrator Markus Baier, whose commitment and enthusiasm for all things IT cannot be appreciated enough. The same holds with respect to administrative issues for Christine Cramer and Christiane Hübner. Thank you all very much!

Finally, I would like to thank my friends and family for their constant support, their loving, patient kindness and the music. Grazie, Ada, per tutto - in particolare per aver sopportato con e per me. Alla fine, tutto *wa behne*.

Darmstadt, October 17, 2022

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

The study of intra- and intercellular biological processes is a formidable challenge. This is in large part because such processes function under high densities of a vast array of functionally distinct and interacting components, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. Each of these components is subject to a multitude of nuisance influences, such as thermal fluctuations, read-off errors from the DNA, misfolding or spontaneous alteration by, e.g., radiation. In addition, their interaction is dependent on tight spatio-temporal coordination which itself is potentially hampered by such nuisances. This shows that in cellular biological systems (as well as biological systems more generally), one usually cannot cleanly isolate the phenomenon of interest and precisely control the boundary conditions of a given experiment. In addition, we are currently only beginning to be able to observe biological processes on the intra- and intercellular level in the wet lab in full detail, as the above conditions make direct and comprehensive observations very challenging.

This yields a setup where experimental observations appear to exhibit a high level of stochasticity. Hence, when analyzing biological experiments, (i) large data sets need to be collected, and (ii) mathematical methods are needed to be able to draw meaningful inferences about underlying systemic effects from these noisy and often sparse data. This thesis deals with such mathematical methods for inference and learning.

The models under study are regime-switching systems: such systems exhibit two or more modes of operation (viz., regimes) that are characterized by qualitatively distinct features. As a first illustrative example, consider a cellular ion channel: ion channels are proteins that are embedded within the cellular membrane and can change their spatial conformation, that is, the relative arrangement of their constituent (sub-)molecules. In one conformation, the channel constitutes

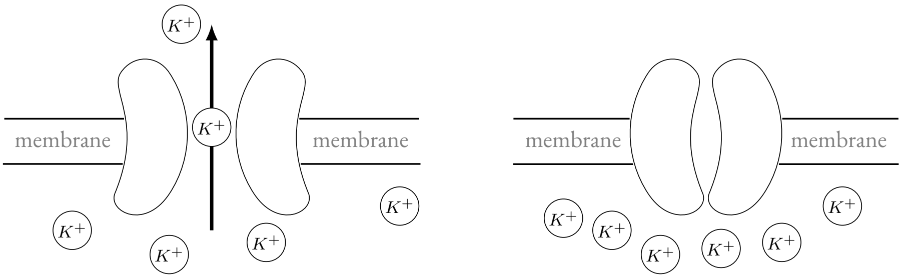


FIGURE 1.1: Conformational switching of ion channels. Left: a potassium channel in the “open” state, allowing a flux of K^+ ions. Right: channel in the “closed” state.

a permeable pore in the membrane through which particular ions can move, see Fig. 1.1. In the other conformation, this pore is closed, such that no ions can pass - it can hence be thought of as an ion gate narrowing and widening depending on the conformation. If the ion current is large enough, this gives rise to an effectively continuous quantity (the current) which depends on the discrete state ("open" vs. "closed") of the channel.

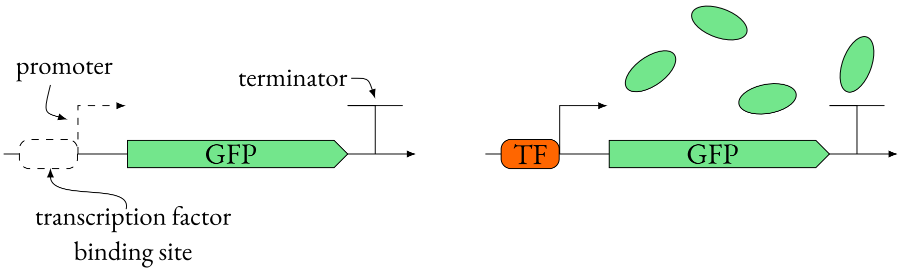
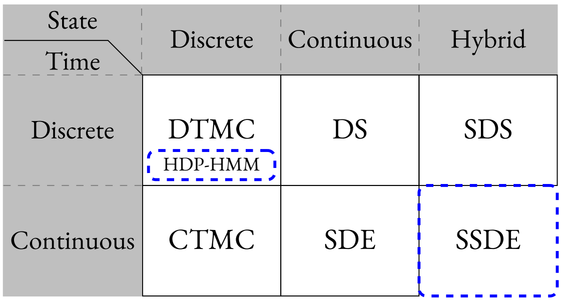


FIGURE 1.2: Illustration of genetic switching. Shown is a strand of DNA with a promoter, gene and terminator. The promoter here is preceded by a transcription factor binding site, which controls the activity of the promoter: only in the presence of a respective transcription factor can the promoter be employed to read off the GFP-coding gene.

The different regimes and the transition between them are often determined by the spatio-temporal context of the system: consider as another example the transcription (TX)-translation (TL) cascade through which proteins are generated from the genetic information stored in the DNA. In the TX phase, an RNA polymerase (RNAP) enzyme binds to a specific portion of the DNA, termed the *promoter*, which sits upstream of the gene of interest (here: green fluorescent protein (GFP), see Fig. 1.2). The RNAP the copies the downstream DNA into a messenger RNA molecule until it reaches a termination signal (the *terminator*). In the second (TL-)phase, the messenger RNAs can be translated into functional proteins by specialized macromolecular machinery. Importantly, in the example shown in Fig. 1.2, the accessibility of the promoter depends on the presence of specific gating molecules: the promoter can only be employed by an RNAP if a transcription factor is bound to the DNA. We are thus dealing with a context-dependent hybrid system structure: the state of the transcription factor binding site (occupied vs. non-occupied) determines the accessibility of the promoter, which - given the presence of RNAPs - modulates the amount of green fluorescent protein (GFP) being generated. As in the ion channel example, this gives rise to a continuous quantity (the GFP fluorescence signal) depending on the discrete occupation state of the transcription factor binding site.

Apart from such hybrid state-space configurations, of course, bio-statistical models can generally also be purely discrete or purely continuous. Additionally, they can be formulated both in discrete and continuous *time*, depending on the experimental setup to be modeled. The possible combinations of space and time structures yield a quite intricate landscape of possible model settings. Relevant to the study of regime switching processes is furthermore the distinction of parametric and nonparametric methods: while the former require an upfront specification of the number of different regimes, the latter enable learning of this number from data. Nonparametric models thus allow more flexibility, but at the cost of increased complexity. Concerning all of these model setups, see Fig. 1.3, an ample literature exists in mathematical statistics, probability theory, and machine learning. The present work fits into this landscape as follows: it is concerned (i) with continuous-time, hybrid state-space switching stochastic differential equation (SSDE) models, and (ii) with discrete-time, discrete state-space hierarchical Dirichlet process hidden Markov



State / Time	Discrete	Continuous	Hybrid
Discrete	DTMC HDP-HMM	DS	SDS
Continuous	CTMC	SDE	SSDE

FIGURE 1.3: Illustration of the methodological scope of this thesis within the existing literature. All model classes - discrete-time Markov chains (DTMCs), dynamical systems (DS), stochastic dynamical systems (SDS), continuous-time Markov chains (CTMCs), stochastic differential equations (SDEs), and switching stochastic differential equations (SSDEs) - are discussed in detail in Chapter 2. In the continuous-time domain, Chapters 3 and 4 build on SSDEs with Markovian switching. In the discrete-time domain, hierarchical Dirichlet process hidden Markov models (HDP-HMMs) - which are a nonparametric extension of DTMCs with noisy observations - are utilized in Chapter 5.

models (HDP-HMMs), where in both cases, the underlying biological processes are assumed to be not observable directly. The utilized process models hence describe *latent* processes, of which only noisy and potentially sparse observations are generated.

In the first setting, the number of distinct regimes is assumed to be known, but the available observational data are sparse and not equidistant in time. Conceptually, this requires the use of continuous-time methods.

In the second setting, the number of qualitatively distinct regimes is unknown. This is often the case when considering conformational dynamics of molecules, e.g. in the analysis of molecular dynamics (MD) simulations or for electrophysiological recordings of ion channel gating. In such settings, data are typically obtained at high frequency and equally spaced in time (as they are either simulated or sampled with high and constant frequency). In this case, we utilize a discrete-time Bayesian nonparametric approach to identify the latent regimes.

1.1 OUTLINE AND CONTRIBUTIONS

The outline of this work is as follows:

CHAPTER 2 provides the mathematical background for the subsequent chapters. It is divided into two parts: first, the (forward) modeling tools will be presented, and second, the (inverse) inference problems of filtering, smoothing and parameter learning will be discussed.

CHAPTER 3 is based on work published in

[1] L. Köhs, B. Alt, and H. Koeppl, "Markov chain Monte Carlo for continuous-time switching dynamical systems", *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, pp. 11 430–11 454, 2022.

A continuous-time, hybrid state-space model is presented for regime-switching biological systems. Concretely, the forward model is formulated via Markov-switching stochastic differential equations. It is shown how to compute the exact posteriors in a Bayesian inference setting. The resulting expressions turn out to be computationally very demanding; to address this challenge, a Gibbs-sampling approximation scheme is then devised. Specifically, all required conditional measures are derived mathematically and it is shown how to draw samples from them. Using synthetic as well as real-world data from an in-house controlled gene switching experiment, the method is shown to perform well both in latent state and parameter inference.

CHAPTER 4 subsequently draws on the results presented in

[2] L. Köhs, B. Alt, and H. Koeppl, "Variational inference for continuous-time switching dynamical systems", *Advances in Neural Information Processing Systems*, vol. 34, 2021.

The sampling-based inference scheme presented in the previous chapter still proves to be computationally expensive. In an effort to decrease the computational burden, we propose a variational approximation to the true posterior. This approximation consists of a set of linear stochastic differential equations between which an underlying Markov jump process switches; this is equivalent to a mixture of Gaussian processes. We argue that this approximation is particularly suited for the analysis of metastable systems, which are of special interest in systems biology and computational biology settings. It is shown that this approach yields accurate state and parameter estimates when the criterion of metastability is fulfilled; the limits of the method are explored by violating metastability. The runtime is found to be drastically improved over the Gibbs sampler.

CHAPTER 5 follows the study

[3] L. Köhs, K. Kukovetz, O. Rauh, and H. Koeppl, "Nonparametric Bayesian inference for meta-stable conformational dynamics", *Physical Biology*, vol. 19, no. 5, p. 056 006, 2022

In this chapter, we focus on a computational structural biology problem, namely the folding of molecules. Here, often only the discrete latent component - corresponding to a set of conformational states - is of interest. Additionally, data are often available on a dense, regular time grid, particularly those stemming from molecular dynamics simulations. These observations are typically real-valued; we utilize a discrete-time, discrete state-space model to approximate this discrete-time, continuous state-space data. A central challenge is the inference of the transition dynamics between conformational states as well as their number, which is typically unknown a priori. This challenge is addressed utilizing a Bayesian nonparametric model, which generalizes the framework of Markov state models (MSMs), an established tool in the field. As the available data sets are typically large, we again pursue a variational inference approach for enhanced scalability compared to existing sampling approaches. This framework yields accurate results both on synthetic as well as real-world benchmark and wet-lab experimental data.

CHAPTER 6 finally provides a summary of the results obtained in this thesis and an outlook sketching potential avenues for further inquiry.

2 BACKGROUND

The problems dealt with in quantitative modeling typically belong to one of two categories: *forward* and *inverse* problems. Forward problems are concerned with predictions of the state of a system given a model f of its dynamics and its associated parameters.

Giving a concrete example building on the process of gene expression (see Fig. 1.2), let $y \in \mathbb{R}_{\geq 0}$ denote the concentration of some protein of interest, e.g., green fluorescent protein (GFP). These proteins are generated at some effective rate $\alpha \in \mathbb{R}_{>0}$ via the processes of transcription and translation as discussed in Chapter 1; at the same time, they are broken down at some other rate $\beta \in \mathbb{R}_{>0}$. This system could be described by an ordinary differential equation (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = \alpha - \beta y(t).$$

By solving this differential equation, one can compute the temporal evolution of the protein concentration and hence answer questions such as "At what time will the GFP level have increased by a factor of more than 100?" or "What will the GFP level be 8 hours after the start of the experiment?".

Knowing the model parameters α, β is however a rare occasion. On the contrary, it is often of interest to establish them from experiments. This is an inverse problem: given a model f and some measurements of the system state $\{y_1, \ldots, y_N\}$ obtained at time points $t_1, \ldots, t_N$, determine its parameters. In realistic settings, the measurements will not correspond to the true state values, but they will be corrupted by some form of measurement error, $\{x_1(y_1), \ldots x_N(y_N)\}$. In the above example, a typical question to be answered could be "What are the production and degradation rates α and β of the studied gene expression system?".

Biological processes such as the production of proteins are hard to study in isolation. It is impossible from a practical perspective to let multiple experiments on the same system start at the same initial conditions. This incomplete knowledge of the initial conditions translates to inherently "noisy" measurements, independent of the accuracy of the used measurement technique: multiple repetitions of the same experiment will yield different results. A statistical mathematical treatment of inference problems is required upfront due to this inevitable uncertainty. Additionally, deterministic models of the unobserved system dynamics (such as the above toy system) are themselves often inadequate to describe many biological phenomena. The above ODE model may describe the average dynamics for high concentrations at best; it is clear that for very low concentrations, a different model would be required to capture the discrete nature of the problem. It also fails to account for the possibility of the population dying out in finite time. When utilizing a discrete model, on the other hand, the production of a particular molecule has to be treated as inherently *stochastic*, as it is impossible to predict, e.g., when an RNAP will bind to the promoter of interest and produce the respective messenger RNA. Therefore, in this as in many systems alike, stochastic models are required to accurately describe the biological processes themselves.

As laid out in the introduction, this thesis is concerned with the modeling of stochastic *switching* biological systems, which are ubiquitously found in biology [4]. Consider as an example again the above gene expression model: the effective production rate α was assumed to be a constant which summarizes all subprocesses necessary to create a functional protein from a stretch of DNA - describing, accordingly, a constant production of proteins. At any given point in time, however, the promoter is either accessible for transcription (if a transcription factor is bound, cf. Fig. 1.2), or it is not, in which case the no messenger RNA can be produced. The above ODE expression model can be extended to reflect these two different states by assuming two different production rates α_1 and α_2, where α_1 represents the "open" state - in which the promoter is accessible to

RNAPs - and α_2 the "closed" one, in which the promoter is blocked: hence, $\alpha_1 > \alpha_2$.[1] With this, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = \alpha_{s(t)} - \beta y(t),$$

with the switching variable $s \in \{1, 2\}$, for which then a respective evolution model has to be defined; the joint system $\{y(t), s(t)\}$ constitutes a continuous-discrete state-space hybrid process. As already argued in Chapter 1, it is often of conceptual interest to identify such separate regimes in which a continous quantity evolves qualitatively different.

Note that additionally, this approach aids computational efficiency: while in principle, RNAPs and proteins are discrete units, their precise number is often secondary. Modeling its detailed evolution via a discrete state-space process would however be computationally very demanding; this computational load can be significantly reduced without loss of important information by approximating the number process by a continuous process representing the stochastic evolution of a concentration value. On the other hand, it is of central interest to know which regime the system is currently in. A hybrid modeling approach hence allows one to make adequate trade-offs between computational tractability and (conceptual) accuracy.

The present chapter provides the background for the subsequent analysis of switching systems. It focuses on some specific classes of linear models, which will be discussed in Section 2.1. The corresponding concepts and methods to address the inverse problem will be laid out in detail in Section 2.2. The purpose of the present chapter is also to serve as a resource and starting point for potential successors. To this end, it tries to strike a balance between rigorosity and accessability. Of course, excellent resources exist on all of the following topics; a wide range of references will be given throughout the chapter.

2.1 STOCHASTIC MODELING OF BIOLOGICAL SYSTEMS

As discussed in the introduction, two central aspects to the mathematical description of stochastic processes are the structure of time and space. For completeness, we first reiterate the definition of a stochastic process [5, 6]: a stochastic process is a set of random variables $\{X(t) : t \in \mathcal{I}\}$ with some index set $\mathcal{I}$. If $\mathcal{I} \subseteq \mathbb{N}$, it is termed a discrete-time process; if $\mathcal{I} \subseteq \mathbb{R}$, it is a continuous-time process. The data one typically obtains from biological experiments are (noisy) observations at discrete time points. As biological processes evolve continuously in time, they are naturally represented in a continuous-time framework. Depending on the temporal structure of the available measurements and the question to be answered, it might however be sufficient to describe the system under study at these discrete time points. Typically, continuous-time descriptions are mathematically more involved, resulting in a trade-off between complexity and expressiveness of the model. Similarly, as stated previously, the state space may be continuous, discrete, or a mixture of both.

This thesis focuses on both discrete- and continuous-time *Markov* processes, that is, processes whose temporal evolution from some time point onwards depends only on the state at that time point, not on its prior history. It discusses both purely discrete state-space systems as well as

[1] Note that of course, more than two different regimes may be introduced, but we deliberately keep this example simple for illustrative purposes.

hybrid processes whose state spaces exhibit discrete and continuous components. This will be formalized for both the discrete-time and the continuous-time case individually: in the following, all necessary background will first be provided for discrete-time models and subsequently for their continuous-time counterparts (cf. Fig. 1.3). For a brief review of the basic concepts of probability theory and the notation used throughout the thesis, please see Appendix A.1.

2.1.1 *Discrete-time Markov models*

First, we consider processes evolving on a fixed time grid $t \in \mathcal{I} \subseteq \mathbb{N}_{>0}$. Notice that throughout, we let discrete-time processes start at $t = 1$ to adhere to conventional notation. A discrete-time process X on some state space $\mathcal{X}$, $X(t) \in \mathcal{X}$, is a Markov process if

$$p(x, t \mid x(t-1), t-1, \ldots, x(1), 1) = p(x, t \mid x(t-1), t-1), \tag{2.1.1}$$

where p denotes the (transition) probability density function (PDF) at time point t, see Appendix A.1 for details. Important results for such models will be discussed below for different state-space configurations.

2.1.1.1 *Discrete-time Markov chains*

A discrete-time Markov chain (DTMC) is a stochastic process $Z := \{Z(t) : t \in \mathcal{I} \subseteq \mathbb{N} > 0\}$ evolving on a discrete-valued state space, $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$, exhibiting the Markov property (2.1.1). Denoting with $\pi : \mathcal{Z} \times \mathcal{Z} \times \mathbb{N}_{\geq 0} \to [0, 1]$ the transition probabilities between all pairs of states $z, z' \in \mathcal{Z}$ for all time points t,

$$\pi(z, z', t) := \mathsf{P}(Z(t+1) = z' \mid Z(t) = z), \tag{2.1.2}$$

the joint density of a full trajectory $Z_{[1,T]} := \{Z(t) : t = 1, \ldots, T\}$ factorizes into a product of such transitions:

$$p(z_{[1,T]}) = p(z(1)) \prod_{t=2}^{T} \pi(z_{t-1}, z_t, t). \tag{2.1.3}$$

A DTMC is hence unambiguously defined via (i) an initial distribution $p(z(1))$, and (ii) the transition function π. The latter can be equivalently expressed as a transition *matrix*,

$$\pi_{zz'}(t) = \pi(z, z', t). \tag{2.1.4}$$

This matrix fulfils $\sum_{z' \in \mathcal{Z}} \pi_{zz'}(t) = 1$ for every time point t. If the transition probabilities are constant over time, $\pi_{zz'}(t) = \pi_{zz'}$, the DTMC is said to be time-homogeneous. As the dynamics of Z are encoded in π, these dynamics can be studied by analyzing its spectral properties. For an accessible and detailed introduction to DTMCs, see, e.g., [7].

2.1.1.2 *Dynamical systems*

Let $\mathcal{Y} \subseteq \mathbb{R}^n$ and $y(t) \in \mathcal{Y}$. A (non-stochastic) dynamical system (DS) is given as

$$y(t+1) = f(y(t), t) \tag{2.1.5}$$

with $f : \mathcal{Y} \times \mathcal{I} \to \mathcal{Y}$ and some initial value $y(1)$ [8]. We will focus on linear dynamical systems (LDS), where

$$y(t+1) = Ay(t) \tag{2.1.6}$$

with $A \in \mathbb{R}^{n \times n}$. As A is constant, Eq. (2.1.6) is a called a time-invariant system, the solution of which can be given explicitly (in the above case simply $y(t) = A^t y(1)$). As for DTMCs, the spectral properties of A encode important information about the process; in particular, the eigenvalues of A determine its stability. If the absolute values of all eigenvalues λ_i of A are smaller than 1, $|\lambda_i| < 1 \, \forall i$, the system converges to a stable state. In the general, time-variant case, this is qualitatively similar, but more involved in the details [8].

A stochastic extension to this LDS is the Gaussian linear dynamical system (GLDS) [9], which is a process $Y := \{Y(t) : t \in \mathcal{I} \subseteq \mathbb{N}_{>0}\}$ on the space $Y(t) \in \mathcal{Y}$:

$$\begin{aligned} Y(t+1) &= AY(t) + BV(t), \\ V(t) &\sim \mathcal{N}(0, \Sigma), \end{aligned} \tag{2.1.7}$$

with $B \in \mathbb{R}^{n \times m}$, $\Sigma \in \mathbb{R}^{m \times m}$, $0 \in \mathbb{R}^m$, and some initial state $Y(1) = y(1)$. Notice that of course, other distributions may be chosen for V, but for the present illustrative purposes, we stick to the simple Gaussian case. In analogy to the non-stochastic case, this system will reach a steady state (in the sense of distributions) if the absolute values of all eigenvalues are smaller than 1 [10].

The joint density of a full trajectory $Y_{[1,T]} := \{Y(t) : t \in \{1, \dots, T\}\}$ of a GLDS can be formulated explicitly as

$$p(y_{[1,T]}) = p(y(1)) \prod_{t=2}^{T} \mathcal{N}(y(t) \mid Ay(t-1), B\Sigma B^\top).$$

Due to the properties of Gaussian distributions, the marginal distribution $p(y, t)$ is also Gaussian at all time points t [9].

2.1.1.3 *Switching dynamical systems*

Combining the above two models for discrete and continuous state spaces, one obtains a stochastic dynamical system (SDS):

$$\begin{aligned} Z(t+1) &\sim \mathsf{P}(Z(t+1) \mid Z(t)), \\ Y(t+1) &= f(Y(t), Z(t+1), t) + B(t+1)V(t), \end{aligned} \tag{2.1.8}$$

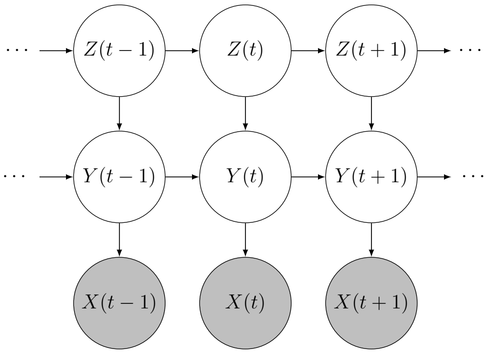


FIGURE 2.1: Graphical model of an switching linear dynamical system (SLDS).

with $f : \mathcal{Y} \times \mathcal{Z} \times \mathbb{R}_{\geq 0} \to \mathcal{Y}$ and V distributed according to some probability measure. Again, we focus on the linear-Gaussian case, that is, switching linear dynamical systems (SLDS) [10, 11]:

$$\begin{aligned} Z(t+1) &\sim \mathsf{P}(Z(t+1) \mid Z(t)), \\ Y(t+1) &= A(Z(t+1))Y(t) + B(Z(t+1))V(t), \\ V(t) &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \end{aligned} \tag{2.1.9}$$

with some initial values $Y(1) = y(1), Z(1) = z(1)$ and Σ, A, B are matrices of appropriate dimensionality; see also Fig. 2.1 The SLDS can be used to model situations in which the system under study has different modes of operation; a globally non-linear behavior can in that way be split into or approximated by locally linear regimes. Due to the switching, the spectral properties of A do not any more alone determine the stability of the system: locally, the linear dynamics may be unstable, but due to frequent switching, the system may still not diverge [12].

The joint density of both trajectories reads

$$\begin{aligned} p(y_{[1,T]}, z_{[1,T]}) = p\left(y(1)\right) p\left(z(1)\right) \\ \cdot \prod_{t=2}^{T} \mathcal{N}(y(t) \mid A(z(t))y(t-1), B(z(t))\Sigma B(z(t))^{\top}) p(z(t) \mid z(t-1)), \end{aligned}$$

showing that the marginal over $y_{[1,T]}$ is not Gaussian anymore, but a complex mixture of Gaussians.

2.1.2 *Continuous-time Markov models*

All of the above discrete-time models have continuous-time analogs. Many of the tools and results from the discrete-time regime transfer to the continuous-time case via a limiting operation on the time grid size; the underlying mathematical concepts however become somewhat more involved.

Consider as in the previous section a general, stochastic process X on some state space $\mathcal{X}$, $X(t) \in \mathcal{X}$, but now let $t \in \mathcal{I} \subseteq \mathbb{R}_{\geq 0}$. Again, we aim to conform to conventional notation and let continuous-time stochastic processes start at $t = 0$, different from the discrete-time case. The process X is a Markov process if

$$p(x, t + \tau \mid x_{[0,t]}) = p(x, t + \tau \mid x(t), t). \tag{2.1.10}$$

This Markov property gives rise to the Chapman-Kolmogorov equation relating the conditional densities at different time points of the process as

$$p(y_1, t_1 \mid y_0, t_0) = \int p(y_1, t_1 \mid y', t')p(y', t' \mid y_0, t_0)\mathrm{d}y'. \tag{2.1.11}$$

The set of all conditional densities hence encodes all information about the process dynamics. Associated with this integral is a differential expression:

$$\mathcal{L}_t f(x) := \lim_{h \to 0} \frac{\mathsf{E}[f(X(t+h)) \mid X(t) = x] - f(x)}{h} \tag{2.1.12}$$

defines the *generator* $\mathcal{L}_t$ of the Markov process, which equivalently characterizes the full process. As an aside, note that Markov processes that can be characterized via such an infinitesimal generator are called *Feller processes*. All processes considered in the context of this thesis fulfil this property. For more background, consult, e.g., [5, 6, 13–15].

2.1.2.1 *Continuous-time Markov chains*

The continuous time analogue to DTMCs are continuous-time Markov chains (CTMCs), which are synonymously called Markov jump processes (MJPs). An MJP Z, with $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$, is fully characterized by (i) an initial probability distribution $p(z(0)) := \mathsf{P}(Z(0) = z(0))\ \forall z(0) \in \mathcal{Z}$, and (ii) the transition rate function defined for $z' \in \mathcal{Z} \setminus z$ as

$$\Lambda(z, z', t) := \lim_{h \searrow 0} \frac{\mathsf{P}(Z(t+h) = z' \mid Z(t) = z)}{h}. \tag{2.1.13}$$

The *exit rate* $\Lambda(z, t) := \Lambda(z, z, t) = -\sum_{z' \in \mathcal{Z} \setminus z} \Lambda(z, z', t)$, which follows from the fact that P is a probability measure. This generally ensures that $\sum_{z' \in \mathcal{Z}} \Lambda(z, z', t) = 0$. In analogy to the discrete-time transition matrix π, the rate function will be abbreviated as $\Lambda_{zz'}(t) := \Lambda(z, z', t)$ in the following, thus aiding concision.

An MJP defines a measure over piece-wise constant paths, cf. Fig. 2.2: concretely, a trajectory $Z_{[0,T]}$ is fully characterized via (i) its *state sequence* $\{Z_k : k = 0, \ldots, K\}$ and (ii) the corresponding *sojourn times* $\{S_k : k = 0, \ldots, K\}$ for which the process resides in these states. The latter also define the *jump times* of the trajectory via

$$J_{k+1} = \sum_{l=0}^{k} S_l,\ \ k \in \{0, \ldots, K\} \tag{2.1.14}$$

and $J_0 := 0$, We denote with $J_{[0,T]} := \{J_0, J_1, \ldots\}$ the set of all jump times occurring in the time interval $[0, T]$. The value of the trajectory at some time point t, $Z(t)$, follows from this via

$$Z(t) = Z_{k^*}, \quad k^* = \max_k \{k : t \geq J_k\}.$$

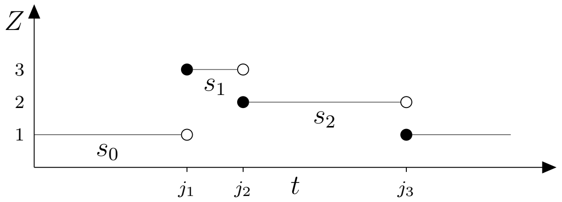


FIGURE 2.2: Sketch of an MJP path. Circles indicate the jumps; the full circles belong to the trajectory and the empty ones do not.

These trajectories are continue à droite, limite à gauche (càdlàg), that is, the value of the trajectory at a jump time t, $Z(t)$, equals the state value *after* the jump occurring at this time.

On the distribution level, the dynamics of this process can be described utilizing the Chapman-Kolmogorov equation (2.1.11). Starting at some initial value $z(s) = z$ evolving towards $z(t) = z'$, $s < t$, Eq. (2.1.11) - with $t' = t - h$ and taking the limit $h \to 0$ - yields the Kolmogorov forward equation (KFE),

$$\frac{\mathrm{d}}{\mathrm{d}t} p(z', t \mid z, s) = \sum_{z'' \in \mathcal{Z}} \Lambda(z'', z', t) p(z'', t \mid z, s), \tag{2.1.15}$$

which for MJPs is called the *master equation*. By inverting this procedure and starting at a terminal value $z(t)$, the corresponding Kolmogorov backward equation (KBE) is derived as

$$\frac{\mathrm{d}}{\mathrm{d}s} p(z', t \mid z, s) = - \sum_{z'' \in \mathcal{Z}} \Lambda(z, z'', t) p(z', t \mid z'', s), \tag{2.1.16}$$

which, analogously, is sometimes called the *backward* master equation.

Any trajectory $Z_{[0,T]}$ consists of an (uncountably) infinite set of points, rendering it impossible to characterize via a product of transition densities akin to Eq. (2.1.3). We can however still explicitly write down a path density, which we formulate for time-homogeneous MJPs for simplicity, $\Lambda(z, z', t) = \Lambda(z, z')$. The definitions of the state sequence and sojourn times provided above allow to explicate

$$\begin{aligned} p(z_{[0,T]}) = \prod_{k=0}^{K} \left\{ \Lambda_{z_k} e^{-\Lambda_{z_k} s_k} \right\}^{\mathbb{1}(z_k = z)} \\ \cdot \left\{ \frac{\Lambda_{z_k z_{k+1}}}{\Lambda_{z_k}} \right\}^{\mathbb{1}(z_k = z \wedge z_{k+1} = z')} \left\{ e^{-\Lambda_{z_K} s_K} \right\}^{\mathbb{1}(z_K = z)}. \end{aligned} \tag{2.1.17}$$

While this likelihood cannot be represented via a product or integral over time, it is possible to provide such an expression for the equivalent of the *ratio* of likelihoods between two MJP measures. Consider an MJP with rate function Λ inducing a measure P and another MJP with rate function $\tilde{\Lambda}$ inducing Q. Then,

$$\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}) = \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\mathsf{Q}} \left(z_{[0,T]} \right) \mathsf{Q}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}) \tag{2.1.18}$$

with the Radon-Nikodym derivative between both measures,

$$\frac{\mathrm{dP}}{\mathrm{dQ}}\left(Z_{[0,T]}\right) = \exp\Bigg\{ -\int_0^T \Lambda(Z(s),s) - \tilde{\Lambda}(Z(s),s)\mathrm{d}s \\ + \sum_{s\in\mathcal{J}_{[0,T]}} \ln\left(\frac{\Lambda(Z(s),s)p(Z(s),s \mid Z(s_-))}{\tilde{\Lambda}(Z(s),s)q(Z(s),s \mid Z(s_-))}\right)\Bigg\}, \quad (2.1.19)$$

see, e.g., [16]. In statistics, this is frequently expressed via expectations over the different measures P and Q, making clear that the Radon-Nikodym itself is a stochastic process [13]:

$$\begin{aligned} \mathsf{E}_{\mathsf{P}}\left[\varphi\left(Z_{[0,T]}\right)\right] &= \mathsf{E}_{\mathsf{Q}}\left[\varphi\left(Z_{[0,T]}\right)\frac{\mathrm{dP}}{\mathrm{dQ}}\left(Z_{[0,T]}\right)\right] \\ &= \int \varphi\left(z_{[0,T]}\right)\frac{\mathrm{dP}}{\mathrm{dQ}}\left(z_{[0,T]}\right) P(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}) \end{aligned}$$

with some test function φ of the stochastic process $Z_{[0,T]}$.

To simulate a time-homogeneous MJP, one can utilize the Doob-Gillespie algorithm [17, 18]: given a current state z,

1. simulate the sojourn time in this state as $S \sim \mathrm{Exp}\left(\Lambda(z)\right)$, and
2. simulate the next state z' via $Z' \sim \mathrm{Cat}\left(\frac{\Lambda(z,1)}{\Lambda(z)}, \ldots \frac{\Lambda(z,z'-1)}{\Lambda(z)}, 0, \frac{\Lambda(z,z'+1)}{\Lambda(z)}, \ldots\right)$.

An extension to time-inhomogeneous MJPs is available with the *thinning* algorithm, which works similarly to the Doob-Gillespie algorithm, but with an additional step to account for changes in the rate function $\Lambda(z, z', t)$ between jumps [19].

2.1.2.2 *Stochastic differential equations*

While discrete-time, non-stochastic linear dynamical systems readily transfer to the continuous-time case by substituting iterative maps such as Eq. (2.1.6) with differential equations, we need to make use of stochastic calculus to define an analog to the stochastic system (2.1.7). For an accessible introduction, see, e.g., [20], for more detailed treatments [5, 6, 15]. The stochastic Itô integral over some function $Q : \mathcal{Y} \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{n\times n}$ is defined as

$$\int_0^t Q(Y(t),t)\mathrm{d}W(t) := \lim_{N\to\infty}\sum_{k=0}^{N} Q(Y(t_k),t_k)\left(W(t_{k+1}) - W(t_k)\right) \quad (2.1.20)$$

on a time grid $\{t_k : k = 1, \ldots, N\}$, $t_0 = 0, t_N = t$, with the n-dimensional Brownian motion $W(t) \in \mathbb{R}^n$. (For completeness, some details on Brownian motion are provided in Appendix A.1.) This allows to define a continuous-time analog to Eq. (2.1.7) as

$$Y(t) = Y(0) + \int_0^t f(Y(t),t)\mathrm{d}t + \int_0^t Q(Y(t),t)\mathrm{d}W(t) \quad (2.1.21)$$

with the drift function $f : \mathcal{Y} \times \mathbb{R}_{\geq 0} \to \mathcal{Y}$ and the invertible dispersion $Q : \mathcal{Y} \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{n \times n}$; this is conventionally abbreviated as a stochastic differential equation (SDE)

$$\mathrm{d}Y(t) = f(Y(t), t)\mathrm{d}t + Q(Y(t), t)\mathrm{d}W(t). \tag{2.1.22}$$

A particularly important result that applies for SDEs is *Itô's lemma* [13]: For a twice differentiable function $\varphi : \mathcal{Y} \times \mathbb{R} \to \mathcal{Y}$,

$$\mathrm{d}\varphi(Y(t), t) = (\partial_t + \mathcal{L}_t)\,(Y(t), t)\mathrm{d}t + \mathrm{d}\mathcal{M}(t) \tag{2.1.23}$$

with the generator of Y,

$$\mathcal{L}_t = \sum_{i=1}^{n} f_i(Y(t), t)\partial_{y_i} + \frac{1}{2}\sum_{i,j=1}^{n} D_{ij}(y, t)\partial_{y_i}\partial_{y_j}, \quad D(y, t) = Q(y, t)Q(y, t)^\top,$$

and the *martingale*

$$\mathrm{d}\mathcal{M}(t) = \sum_{i,j=1}^{n} \partial_{y_i}\varphi(Y(t), t)D_{ij}(Y(t), t)\mathrm{d}W_j(t).$$

Martingales are stochastic processes for which the conditional expectation given the history of the process is equal to the present value:

$$\mathsf{E}\big[\mathcal{M}(t') \mid \mathcal{M}_{[0,t]}\big] = \mathcal{M}(t)$$

for $t' > t$. In the context of this thesis, only standard Brownian motion martingales occur, the conditional expectations of which by definition equate to zero. Note, however, that martingales are a richly-structured, extensive field of research; for an introduction, see, e.g., [6]. Itô's lemma applies not only to SDEs, but more generally to semimartingales, which are martingales that can be decomposed into a martingale and a càdlàg process [5, 6]; in particular, it also holds for MJPs [13].

From Itô's lemma, the KBE and KFE for SDEs can be derived; this derivation is provided in Appendix A.1. The KBE prescribes the evolution of the conditional density at some time point $s < t$ with $y(s) = y, y(t) = y'$ and is obtained as

$$\begin{aligned}\partial_s p(y', t \mid y, s) = {} & \sum_{i=1}^{n} f_i(y, t)\partial_{y_i} p(y', t \mid y, s) \\ & + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(y', t)\partial_{y_i}\partial_{y_j} p(y', t \mid y, s) = -\mathcal{L}_t p(y', t \mid y, s).\end{aligned} \tag{2.1.24}$$

The Kolmogorov forward equation for SDEs is also known as Fokker-Planck equation (FPE) and is obtained as the adjoint

$$\begin{aligned}\partial_t p(y', t \mid y, s) = {} & -\sum_{i=1}^{n} \partial_{y'_i} f_i(y', t) p(y', t \mid y, s) \\ & + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y'_i}\partial_{y'_j} D_{ij}(y', t) p(y', t \mid y, s) = \mathcal{L}_t^\dagger p(y', t \mid y, s).\end{aligned} \tag{2.1.25}$$

Notice the conceptual similarity to the MJP case, cf. Eqs. (2.1.15) and (2.1.16); this is unsurprising, as both MJPs and SDEs are examples of Feller processes as mentioned above.

For linear functions f and fixed $Y(0) = y(0)$, the FPE yields time-dependent Gaussians as conditional marginals [21]; it is however not possible - as it still was for the MJP - to write down a density function over trajectories $y_{[0,T]}$. Intuitively, this would need to be something like an "infinite product of Gaussians" (cf. [20]), which cannot be rigorously defined. It is, however, still possible to compute the Radon-Nikodym derivative between two measures induced by SDEs, akin to Eq. (2.1.19): consider two SDEs driven by different Brownian motions W_{P} and W_{Q}, inducing different measures P and Q:

$$\begin{aligned} \mathrm{d}Y(t) &= f(Y(t),t)\mathrm{d}t + Q(Y(t),t)\mathrm{d}W_{\mathsf{P}}(t) \quad &\longleftrightarrow \quad &\mathsf{P}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]}), \\ \mathrm{d}\tilde{Y}(t) &= g(\tilde{Y}(t),t)\mathrm{d}t + Q(\tilde{Y}(t),t)\mathrm{d}W_{\mathsf{Q}}(t) \quad &\longleftrightarrow \quad &\mathsf{Q}(\tilde{Y}_{[0,T]} \in \mathrm{d}y_{[0,T]}). \end{aligned}$$

The Radon-Nikodym derivative between both measures can be computed as [22, 23]

$$\begin{aligned} \frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}}\left(Y_{[0,T]}\right) = \exp\Bigg\{ &\int_0^T \left(f(Y(t),t) - g(Y(t),t)\right)^\top D(Y(t),t)^{-1}\mathrm{d}W_{\mathsf{P}}(t) \\ -\frac{1}{2}\int_0^T &\left(f(Y(t),t) - g(Y(t),t)\right)^\top D(Y(t),t)^{-1}\left(f(Y(t),t) - g(Y(t),t)\right)\mathrm{d}t \Bigg\} \end{aligned} \tag{2.1.26}$$

with D as before; see the also discussion regarding Eq. (2.1.19). As an aside, note that this can even be generalized to differing dispersion matrices [24]. Defining the difference process

$$\xi(t) := Q(Y(t),t)^{-1}f(Y(t),t) - g(Y(t),t),$$

both measures are related via Girsanov's theorem [15], which tells us that

$$W_{\mathsf{Q}}(t) = W_{\mathsf{P}}(t) - \xi(t) \tag{2.1.27}$$

is a Brownian motion under Q, meaning that $\mathsf{E}_{\mathsf{Q}}\left[W_{\mathsf{Q}}\right] = 0$, cf. the above remarks about martingales.

To simulate SDEs, serveral numerical schemes are available [25]. The most straightforward approach is the Euler-Maruyama approximation: here, the time domain is discretized with some step size h, $\{0 = t_0, t_1 = t_0 + h, \dots t_N = T\}$, and the process approximation is given via

$$Y(t_k) = Y(t_{k-1}) + f\left(Y(t_{k-1}), t_{k-1}\right)h + Q\left(Y(t_{k-1}), t_{k-1}\right)\Delta W(t_{k-1}). \tag{2.1.28}$$

The random variables

$$\Delta W(t_{k-1}) := W(t_k) - W(t_{k-1})$$

are normally distributed with mean zero and variance h, as prescribed by the definition of Brownian motion.

2.1.2.3 *Switching stochastic differential equations*

Analogous to SLDS, cf. Fig. 2.1, we can combine MJPs and SDEs to a discrete-continuous hybrid systems. The MJP-switching stochastic differential equation (SSDE) is composed as follows: the switching process Z at the top of the hierarchy is given as an MJP. The subordinated diffusion process Y is a continuous-valued process, $Y(t) \in \mathcal{Y} \subseteq \mathbb{R}^n$, depending on the freely evolving MJP Z. It is defined as an SDE conditional on the state of the MJP Z, yielding an SSDE [26]:

$$\mathrm{d}Y(t) = f(Y(t), Z(t), t)\,\mathrm{d}t + Q(Y(t), Z(t), t)\,\mathrm{d}W(t), \tag{2.1.29}$$

with the drift function $f : \mathcal{Y} \times \mathcal{Z} \times \mathbb{R}_{\geq 0} \to \mathcal{Y}$ and the invertible dispersion $Q : \mathcal{Y} \times \mathcal{Z} \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{n \times n}$ determining the noise covariance as $D(y, z, t) := Q(y, z, t) Q^\top(y, z, t)$. Given a realization of the MJP, the SSDE in Eq. (2.1.29) can hence be equivalently interpreted as a concatenation of individual SDEs determined by the MJP. Note that throughout this thesis, the terms "hybrid process" and "hybrid system" will refer to MJP-SSDEs unless stated otherwise.

2.2 BAYESIAN INFERENCE FOR BIOLOGICAL SYSTEMS

Having defined a model for the system of interest, we want to utilize it for inference: statistical inference refers to the process of drawing conclusions about properties θ of a statistical model based on some given data set of observations x. The centerpiece of *Bayesian* inference is Bayes' theorem,

$$p(\theta \mid x) = \frac{p(x \mid \theta)}{p(x)} p(\theta), \tag{2.2.30}$$

which yields a *posterior distribution* over the desired quantity θ in light of the observed data. The term $p(x \mid \theta)$ denotes the likelihood function, $p(x)$ the (model) evidence and $p(\theta)$ the prior distribution. It is sometimes (in the machine learning literature in particular) differentiated between inference and learning, where inference refers to the estimation of latent state variables and learning to the estimation of parameters. In fully Bayesian settings, this distinction is not as meaningful, since both the latent processes and the system parameters are treated as random variables and Eq. (2.2.30) applies similarly to either one. For clarity, however, we will keep with this distinction.

The observed data which are used for inference are typically recorded at discrete time points, in particular in biological experiments (but also in engineering contexts, see, e.g., [27]). If we model the latent processes in continuous time, this defines a *continuous-discrete* inference problem: we aim to draw inferences about a latent continuous-time process from only a finite set of observations. This problem class has a long history in the filtering community [20, 28, 29].

In the following, we keep with the notational conventions laid out in the previous section: discrete-time processes start at $t = 1$, continuous-time processes at $t = 0$. A discrete-time observation trajectory in the time interval $[1, T] \subset \mathbb{N}$ is correspondingly denoted as $x_{[1,T]} := \{x(1), \dots, x(T)\}$.[2] In the continuous-discrete setting, the observational data obtained in the

[2] Notice that this plausibilizes the different starting time points in the discrete- and continuous-time settings: if one had chosen $t = 0$ as a starting point in the discrete regime, the set $x_{[0,T]}$ would contain $T + 1$ observations, which is somewhat unwieldy. In the continuous-time case, on the other hand, the interval $[0, T]$ is of length T, as one would expect.

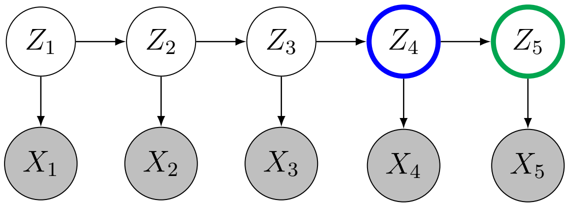


FIGURE 2.3: Illustration of the filtering and the smoothing problems for a discrete-time HMM. The filtering problem consists in determining the marginal probability $p(z,t \mid x_{[1,t]})$ of the latent state at some time point t given observations up to and including time t (green circle); the smoothing problem consists in finding $p(z,t \mid x_{[1,T]})$ given also "future" observations at times $T > t$ (blue circle).

interval $[0,T] \subset \mathbb{R}$ is denoted as $x_{[0,T]} := \{x(t_1) = x_1, \ldots, x(t_N) = x_N\}$ with observation time points $0 \leq t_1, \ldots, t_N \leq T$.

2.2.1 *Filtering and smoothing problems*

In inference problems, one is typically interested in the best estimate of (a function of) the latent state at a given point in time. This gives rise to the filtering problem and the smoothing problem. Their general structure is illustrated with the classic hidden Markov model (HMM), see Fig. 2.3: an HMM is defined via a latent DTMC $Z = \{Z(t) : t \in \{1, \ldots, T\}\}$ as described in Section 2.1.1.1, giving rise to observations $\{X(t) : t \in \{1, \ldots, T\}\}$. An observation at time point t, $X(t)$, depends only on the state of the latent process at the same time t. This dependency is given by the observation density

$$p(x(t) \mid z(t), \{\theta_1, \ldots, \theta_{|\mathcal{Z}|}\}) = p(x(t) \mid \theta_{z(t)}), \tag{2.2.31}$$

where $\{\theta_i : i = 1, \ldots, |\mathcal{Z}|\}$ represents a set of generic distribution parameters for each state i. The *filtering problem* consists in finding the marginal distribution of the latent process at some time t given the observational data up to and including time point t:

$$\mathsf{P}(Z(t) = z \mid X(1) = x(1), \ldots, X(t) = x(t)).$$

Formulated more generally to incude the continuous-time case, we are interested in

$$\mathsf{E}\left[\varphi(Z(t)) \mid x_{[0,t_i]}\right] \tag{2.2.32}$$

for some function φ of the latent state, and $t_i := \max\{t' : t' \in \{t_1, \ldots, t_N\}, t' \leq t\}$.

The *smoothing problem* consists in finding the marginal distribution of the latent process given also observations from future time points:

$$\mathsf{P}(Z(t) = z \mid X(1) = x(1), \ldots, X(T) = x(T)).$$

More generally, we aim to compute

$$\mathsf{E}\left[\varphi(Z(t)) \mid x_{[0,T]}\right]. \tag{2.2.33}$$

The smoothing problem hence incorporates *more* observed information into the estimate, typically resulting in "smoother" estimates of the latent state - hence the name. In the following, we will be concerned with this latter problem. For more in-depth information about stochastic filtering and smoothing, see, e.g., [30] and [31].

2.2.1.1 *Discrete time*

To efficiently compute the time point-wise smoothing marginals of discrete-time processes, Bayes' rule can be employed. For the latent process Z of the above-defined HMM, this allows to straightforwardly derive a forward-backward message-passing algorithm [9]: the desired posterior marginal density at time point t can be written as

$$p(z,t \mid x_{[1,T]}) \propto \alpha(z,t)\beta(z,t) \tag{2.2.34}$$

with the forward-messages

$$\alpha(z,t) := p(z,t \mid x(1),...,x(t))$$

and the backward-messages

$$\beta(z,t) := p(x(t+1),...,x(T) \mid z,t).$$

By applying the law of total probability and the conditional independence relations encoded in the graphical model (see Fig. 2.3), one can by straightforward marginalization derive recursions for both α and β as

$$\begin{aligned}\alpha(z,t) &= p(x(t) \mid z) \sum_{z' \in \mathcal{S}} \alpha(z',t-1)\pi_{z'z},\\ \beta(z,t) &= \sum_{z' \in \mathcal{S}} \beta(z',t+1)p(x(t+1) \mid z')\pi_{zz'}.\end{aligned} \tag{2.2.35}$$

A similar algorithm known as the Rauch-Tung-Striebel (RTS) smoother (which builds on the famous Kalman filter [21]) exists for LDS.

2.2.1.2 *Continuous time*

In the continuous-time case, one has to differentiate between continuous-discrete models in which observations are obtained only at individual time points (see the discussion in the beginning of Section 2.2) and fully continuous models, where complete trajectories are observed. For continuous-discrete models, it is possible (similar to the discrete-time case) to apply Bayes' rule, as the number of observations is finite: for the continuous-time HMM, for instance, this yields

$$p(z,t \mid x_{[0,T]}) \propto \alpha(z,t)\beta(z,t), \tag{2.2.36}$$

with, in analogy to the discrete-time case, the (forward) filtering density

$$\alpha(z,t) := p(z,t \mid x_{[0,t_k]})$$

and the (backward) function

$$\beta(z,t) := p(x_{[t_{k+1},T]} \mid z,t),$$

where $t_k := \max\{t' \in \{t_1,\ldots,t_N\} : t' \leq t\}$. Instead of a recursive map akin to Eq. (2.2.35), one can derive forward and backward ODEs describing the dynamics of these quantities [32].

If, on the other hand, $x_{[0,T]} := \{x(t) : t \in [0,T] \subset \mathbb{R}_{\geq 0}\}$, i.e., the observations consist of a full continuous trajectory, one cannot directly replicate these results, as it is not possible to apply the conventional Bayes rule. There is however a continuous-time analog called the *Kallianpur-Striebel* formula [13]:

$$\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid x_{[0,T]}) = \frac{G(x_{[0,T]}, z_{[0,T]})\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]})}{\int G(x_{[0,T]}, z'_{[0,T]})\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z'_{[0,T]})}, \tag{2.2.37}$$

where we define

$$G(x_{[0,T]}, z_{[0,T]}) := \frac{\mathsf{P}(X_{[0,T]} \in \mathrm{d}x_{[0,T]} \mid z_{[0,T]})}{\mathsf{P}(X_{[0,T]} \in \mathrm{d}x_{[0,T]})}$$

as a shortcut for the Radon-Nikodym derivative between measures. This allows to derive (stochastic) partial differential equations (PDEs) describing the forward and backward dynamics of the posterior process. A particularly well-known result is the Wonham filter, which is obtained for a latent MJP with an SDE observation process. Notice that similar results are also obtained for different combinations of latent and observed process classes, see, e.g., [31].

2.2.2 *Conjugate priors*

The prior distribution $p(\theta)$ characteristic for Bayesian inference has to be chosen by hand. It may, for instance, be set based on knowledge from previous experiments. With respect to the computational tractability of the inference problem it is beneficial to choose priors that are *conjugate* to the respective likelihoods: a prior probability density parameterized by γ, $p(\theta \mid \gamma)$, is said to be conjugate to a likelihood density $p(x \mid \theta)$ if the resulting Bayesian posterior is of the same functional form as the prior,

$$p(x \mid \theta)p(\theta \mid \gamma) \propto p(\theta \mid \gamma'). \tag{2.2.38}$$

The updated parameter γ' is a function of the prior parameter and the data, $\gamma' = \gamma'(\gamma, x)$. This property greatly simplifies inference because the computation of a posterior distributions then reduces to computing the parameter update $\gamma \rightarrow \gamma'$. As an aside, note that this can also be generalized to distributions that do not admit probability densities [33].

The parameter updates become particularly simple if *exponential family* priors are utilized: let $\theta \in \mathcal{A}$, $\gamma \in \mathcal{B}$. The class of distributions that admit a density of the form

$$p(\theta \mid \gamma) = h(\theta) \exp\left\{\langle \eta(\gamma), T(\theta)\rangle - \ln g(\eta)\right\} \tag{2.2.39}$$

is called the exponential family. The function $h : \mathcal{A} \rightarrow \mathbb{R}_{\geq 0}$ is called the *base measure*, $\eta : \mathcal{B} \rightarrow \mathcal{H}$ the *natural parameter* and $T : \mathcal{A} \rightarrow \mathcal{H}$ the *sufficient statistic*, where $\mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle$: $\mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$. The function $g : \mathcal{H} \rightarrow \mathbb{R}_{>0}$ ensures normalization. For the purposes of this thesis, it is sufficient to consider $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}^n$. It can be shown that for every exponential family distribution, a conjugate prior exists [9], which can be written as

$$p(\eta \mid \chi, \nu) = f(\chi, \nu) \exp\left\{\nu\langle \eta, \chi\rangle + \nu \ln g(\eta)\right\}. \tag{2.2.40}$$

with g as above, $\chi \in \mathcal{H}$, $\nu \in \mathbb{R}$ and a normalizing function $f : \mathcal{H} \times \mathbb{R} \rightarrow \mathbb{R}$.

EXAMPLE: DIRICHLET-MULTINOMIAL CONJUGACY As a concrete example with distributions that will be used in Chapter 5, consider the multinomial distribution

$$\mathrm{Mult}(n \mid \pi) = \frac{\Gamma\left(\sum_i n_i + 1\right)}{\prod_i \Gamma(n_i + 1)} \prod_{i=1}^{K} \pi_i^{n_i}, \tag{2.2.41}$$

with $\pi := \{\pi_1, ..., \pi_K\} \in \Delta^K$ denoting the success probabilities of the K categories, $n := \{n_1, ..., n_K\} \in \mathbb{N}^K$ and $n_i \in \mathbb{N}$ denoting the number of occurrences of category i. The number of total events $\sum_i n_i =: N$. This distribution is a member of the exponential family, as can be seen by identifying

$$\begin{aligned} \eta(\pi) &= (\ln \pi_1, ..., \ln \pi_K)^\top, & T(n) &= (n_1, ..., n_K)^\top, \\ h(n) &= \frac{N!}{\prod_i n_i!} = \frac{\Gamma\left(\sum_i n_i + 1\right)}{\prod_i \Gamma(n_i + 1)}, & g(\pi) &= 1. \end{aligned}$$

The conjugate prior to the multinomial is the Dirichlet distribution

$$\mathrm{Dir}(\pi \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \pi_i^{\alpha_i - 1} \tag{2.2.42}$$

with $\alpha := \{\alpha_1, ..., \alpha_K\}$, $\alpha_i \in \mathbb{R}_{>0}$; compare this to Eq. (2.2.40) with the identities

$$\begin{aligned} \nu &= 1, \\ \chi &= (\alpha_1 - 1, ..., \alpha_K - 1)^\top, \\ f(\chi, \nu) &= \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)}. \end{aligned}$$

Straightforward multiplication of the likelihood Eq. (2.2.39) with the prior Eq. (2.2.40) yields the generic update rule $\gamma'(\gamma, x)$:

$$\nu' = \nu + 1, \quad \chi' = \chi + T(\theta).$$

For the Dirichlet-multinomial conjugacy between Eqs. (2.2.41) and (2.2.42), this translates to

$$\alpha_i' = \alpha_i + n_i.$$

For a comprehensive overview over distributions and conjugate priors, see, e.g., [34].

2.2.3 *Nonparametric models*

A particular challenge regarding the specification of a suitable prior distribution that will arise in Chapter 5 is the specification of a prior over countably infinite states. This is enabled by the *Dirichlet process (DP)*: a DP is a stochastic process taking values in the space of (discrete) probability measures [35]. Let $\gamma \in \mathbb{R}_{>0}$ and P_0 be a probability measure on some measurable space $(\mathcal{S}, \mathcal{F})$. The stochastic process $\mathrm{DP}(\gamma, \mathsf{P}_0)$ is called a Dirichlet process if for any finite partition $\mathcal{A} := A_1, ..., A_K$ of $\mathcal{S}$, i.e., $\bigcup_i A_i = \mathcal{S}$ and $A_i \cap A_j = \emptyset \; \forall \, i \neq j$, the random variable $\mathsf{P}_1 \sim \mathrm{DP}(\gamma, \mathsf{P}_0)$ is Dirichlet distributed on this partition, [36]

$$\mathsf{P}_1 \sim \mathrm{DP}(\gamma, \mathsf{P}_0) \iff (\mathsf{P}_1(A_1), ..., \mathsf{P}_1(A_K)) \sim \mathrm{Dir}\left(\gamma \, \mathsf{P}_0(A_1), ..., \gamma \, \mathsf{P}_0(A_K)\right). \quad (2.2.43)$$

P_1 hence is a random measure on $(\mathcal{S}, \mathcal{F})$; $\int \mathrm{d}\mathsf{P}_1 = 1$ almost surely. The parameter γ prescribes the concentration of the process: in the limit $\gamma \to 0$, all realizations of the process would be bound to one single point. The *base measure* P_0 determines the expected value of the DP: $\mathsf{E}[\mathsf{P}_1] = \mathsf{P}_0$.

Importantly, the conjugacy between the multinomial and the Dirichlet distribution discussed in the previous section carries over to the countably infinite case: given a set of samples from the measure P_1, $\theta_i \sim \mathsf{P}_1$ for $i = 1, ..., N$, the posterior distribution of P_1 given these data is again DP-distributed [36]:

$$\mathsf{P}_1 \mid \{\theta_i : i = 1, ..., N\} \sim \mathrm{DP}\left(\gamma + N, \frac{\gamma}{\gamma + N}\, \mathsf{P}_0 + \frac{1}{\gamma + N} \sum_{i=1}^{N} \delta_{\theta_i}\right). \quad (2.2.44)$$

Due to this result, the definition Eq. (2.2.43) is helpful in inference, as will be seen in Chapter 5. It is however non-constructive, i.e., it provides no method to actually draw samples from a DP, which would be required for the construction of a generative model.

An equivalent, constructive definition of the DP can be formulated as follows [37]. To generate a DP-distributed probability measure $\mathsf{P}_1 \sim \mathrm{DP}(\gamma, \mathsf{P}_0)$, it takes i.i.d. random variables distributed according to the base measure,

$$\Theta_k \overset{\text{i.i.d.}}{\sim} \mathsf{P}_0 \text{ for } k = 1, 2, ... \quad (2.2.45)$$

To each Θ_k then a probability mass B_k is assigned via the so-called *stick-breaking process*:

$$\begin{aligned} \epsilon_k &\overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \gamma) \text{ for } k = 1, 2, ..., \\ B_k &= \epsilon_k \prod_{j=1}^{k-1} (1 - \epsilon_j). \end{aligned} \quad (2.2.46)$$

Eq. (2.2.46) is compactly written as $B \sim \mathrm{GEM}(\gamma)$ (as an acronym for the names Griffiths, Engen and McCloskey) [38]. For a visual representation of this process for a realization $B = \beta$, see Fig. 2.4. By construction, $\sum_k B_k = 1$ almost surely. The random measure $\mathsf{P}_1 \sim \mathrm{DP}(\gamma, \mathsf{P}_0)$ then reads

$$\mathsf{P}_1 = \sum_{k=1}^{\infty} B_k \delta_{\Theta_k}. \quad (2.2.47)$$

Note that the stick-breaking measure B encourages sparsity by allocating decreasing amounts of probability mass to higher state indices k.

2.2.4 *Approximation methods*

The desired posterior distributions (or expectations, more generally) typically are complicated objects that do not allow analytical solutions. As detailed above, in some cases, exact solutions of

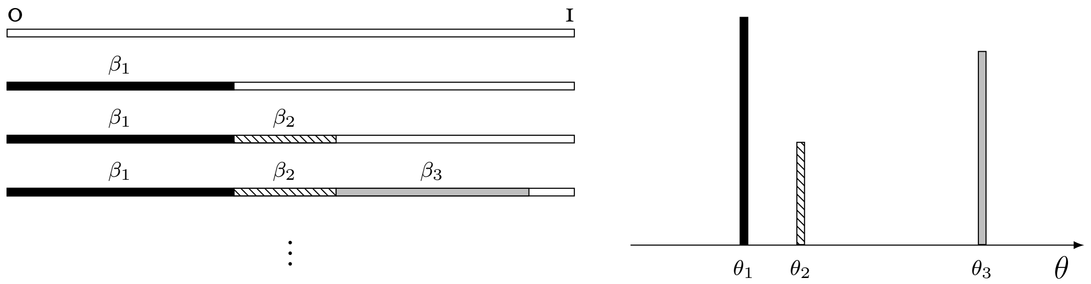


FIGURE 2.4: Sketch of the stick-breaking process Eq. (2.2.46). A "stick" of length 1 is successively broken into pieces: first, a piece of length $\epsilon_1 \sim \mathrm{Beta}(1, \gamma)$ is taken off the stick and $\epsilon_1 =: \beta_1$. The remainder of the stick accordingly has length $(1 - \epsilon_1)$. Of this remainder, again a portion $\epsilon_2 \sim \mathrm{Beta}(1, \gamma)$ is taken off; this then has length $\beta_2 := \epsilon_2(1 - \epsilon_1)$. This procedure repeats indefinitely and is guaranteed to sum up to one, $\sum_i \beta_i = 1$. In the DP, the stick lengths β_i serve as respective probability masses of the atoms θ_i drawn from the base measure P_0.

the respective dynamics (message passing algorithms, ODEs or PDEs) can be found [39] - this is not generally the case, however, and even if it is, one then is often confronted with the problem of computational tractability, e.g., by the curse of dimensionality. Consequentially, a great wealth of approximation techniques has been developed, which can be employed to devise tractable algorithms [40–43]. In the following, two main approaches for approximate inference will be presented that are relevant to the present thesis, namely, Markov chain Monte Carlo and variational inference.

2.2.4.1 *Markov chain Monte Carlo*

Markov chain Monte Carlo (MCMC) methods constitute a powerful class of algorithms allowing one to efficiently sample from complex probability distributions. This is achieved by defining a Markov chain which can be tractably propagated and which has the desired distribution as its stationary distribution. MCMC methods have been subject to intense study for decades; accordingly, a wide array of methods exist [44–46]. Of particular interest to this thesis are Gibbs-type sampling schemes, which enable sampling from a complex joint distribution by only knowing the full conditionals: let $x \in \mathbb{R}^n$ and x_k^i denote the k-th component of the i-th sample. Choose some initial x^1 and then sample successively from

$$X_k^i \sim p(x_k^i \mid x_1^i, \ldots, x_{k-1}^i, x_{k+1}^{i-1}, \ldots, x_n^{i-1}) \tag{2.2.48}$$

over all components k and as many iterations I as desired. Provided that the distribution is strictly positive, $p(x) > 0$, this generates I samples from the joint distribution $p(x_1, \ldots, x_n)$. Note that without the positivity condition, the latter is *not* true, as then the set of full conditionals does not necessarily amount to a valid probability distribution; see, e.g., [47] and [48]. While MCMC approaches are very flexible as they allow also complex, high-dimensional distributions to be sampled, they produce correlated samples which depend on the initialization of the algorithm. Consequently, one needs to discard many samples throughout in order to reduce correlations (in an effort to obtain i.i.d. samples; *thinning*) and additionally some samples from the beginning (to reduce dependency on the initialization, *burn-in*). For Monte Carlo approaches to discrete-time as well as continuous-time systems, see, e.g., [40, 41, 49].

2.2.4.2 *Variational inference*

The main problem with MCMC methods is the associated computational cost in the face of large data sets or complex models. An alternative approach that addresses this issue is variational inference (VI), which casts the inference task as an optimization problem [50]. VI methods are widely used, see, e.g., [51–54]; their goal is to identify the best approxmation to the exact (intractable) posterior measure over the process of interest Z, $\mathsf{P}_x := \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid x_{[0,T]})$ by an auxiliary measure $\mathsf{Q} := \mathsf{Q}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]})$,

$$\mathsf{Q}^* := \arg\min_{\mathsf{Q}} \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x]. \tag{2.2.49}$$

The cost function is given as the Kullback-Leibler (KL) divergence

$$\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x] = \mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}_x}\right], \tag{2.2.50}$$

where the expectation is taken with respect to Q. Considering the KL divergence between the Q-measure and the prior measure $\mathsf{P} := \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]})$, notice that [23]

$$\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}] = \mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}_x} + \ln \frac{\mathrm{d}\mathsf{P}_x}{\mathrm{d}\mathsf{P}}\right].$$

Observing that

$$\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid x_{[0,T]}) \propto \mathsf{P}(X_{[0,T]} \in \mathrm{d}x_{[0,T]} \mid z_{[0,T]})\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]})$$

and utilizing the Kallianpur-Striebel formula (2.2.37), we find

$$\frac{\mathrm{d}\mathsf{P}_x}{\mathrm{d}\mathsf{P}} = \frac{\prod_{i=1}^{N} p(x(t_i) \mid z(t_i))}{p(x_{[0,T]})}, \tag{2.2.51}$$

which we can write in terms of densities assuming that the observation process $X \mid Z$ admits such a density. This allows to reformulate the KL divergence (2.2.50) for general continuous-discrete processes as

$$\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x] = \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}] - \sum_{i=1}^{N} \mathsf{E}[\ln p(x(t_i) \mid z(t_i))] + \ln p(x_{[0,T]}). \tag{2.2.52}$$

This result holds analogously for discrete-time processes. Notice that a similar decomposition can be obtained also more generally for continuously observed processes, see, e.g., [23, 55]. For more details on the KL divergence between stochastic processes, see [56, 57]. As the KL divergence is bounded by zero, $\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}] \geq 0$, the minimum is attained at zero for the exact (and intractable) posterior $\mathsf{Q} = \mathsf{P}_x$. This is reflected also in the KL decomposition, as the right-hand side contains the marginal likelihood $p(x_{[0,T]})$, which is the problematic part with respect to computability. Due to the boundedness, the above can, however, be reformulated as an inequality reading [58]

$$\ln p(x_{[0,T]}) \geq \sum_{i=1}^{N} \mathsf{E}[p(x(t_i) \mid z(t_i))] - \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}] =: \mathsf{L}.$$

The so-defined quantity L is the evidence lower-bound (ELBO); as $\ln p(x_{[0,T]})$ does not depend on Q, *minimization* of the KL (2.2.50) is equivalent to *maximization* of the ELBO:

$$\arg\min_{\mathsf{Q}} \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x] \iff \arg\max_{\mathsf{Q}} \mathsf{L}. \tag{2.2.53}$$

To be able to optimize the ELBO efficiently, further assumptions typically have to be made regarding Q, such as the classic mean-field assumption:

$$\mathsf{Q}\left((Z_{[0,T]}, \Theta) \in \mathrm{d}(z_{[0,T]}, \Theta)\right) = \mathsf{Q}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]})\mathsf{Q}(\Theta \in \mathrm{d}\theta).$$

In the discrete-time case with conjugate model priors and likelihoods, this yields the coordinate-ascent variational inference (CAVI) algorithm, enabling one to efficiently update one mean-field component at a time, see, e.g., [9, 50] and Chapter 5. If further constraints are placed on Q, such as a positive semi-definiteness constraint on Gaussian covariances, the now constrained optimization problem can be converted into an unconstrained problem by means of Lagrangian multipliers [9]. In the continuous-time case, this holds similarly with Lagrange multiplier *functions*. For a thorough introduction to optimization and control theory, see [59].

3 MARKOV CHAIN MONTE CARLO FOR HYBRID SYSTEMS

As laid out previously, a wide range of mathematical tools exists to model the dynamics of biological systems. The framework of choice naturally is prescribed by the system under study and the scientific question to be answered. For instance, when modeling the conformational gating behavior of an ion channel (see the example given in Chapter 1), one is typically interested in the succession of "open" and "closed" states, permitting or preventing ions from flowing through; as the random thermal motion of the constituent atoms of the channel molecule around the current conformational state does not provide relevant information, these can be safely abstracted away by modeling the switching via a DTMC or MJP on a purely discrete state space $\mathcal{Z} \subseteq \mathbb{N}$.

In many systems, however, both discrete *and* continuous components are of interest, meaning that neither can be neglected upfront; two respective examples were already discussed in Chapter 1. Going beyond these examples, systems with a hybrid state-space structure are found over a wide range of natural and engineering sciences: in neuroscience, for instance, the brain is commonly assumed to adopt different states depending, e.g., on the environment or own actions, such as "eyes opened" versus "eyes closed", eliciting qualitatively distinct continuous electrophysiological dynamics [60]. These different states and the transition dynamics between them may influence, e.g., decision-making and cognitive performance [61]. Different dynamical regimes are indeed found on multiple time and lengths scales using a variety of experimental techniques [62]. In addition to aiding understanding, probabilistic switching models open up the possibility of identifying appropriate external control stimuli that trigger transitions between states, e.g., from sleep to wakefulness [63].

A similarly wide array of examples may also be found in cellular biology: for instance, the - discrete - states of genetic toggle switches drive continuous measurable quantities such as protein concentrations [64]. As discussed in Chapter 1, one may not only be interested in the conformational state of an ion channel, but also in the current passing through [65], requiring hybrid models for its gating behavior. Hybrid system frameworks are also used to model DNA replication [66, 67] and phenotype differentiation in systems biology [68].

Examples from branches of engineering sciences include safety modeling of air traffic [69] or electrical power systems [70] under potential system failures; more generally, hybrid systems are - somewhat akin to the neuroscience examples - used in electrical engineering to control and optimize the power distribution in an electrical grid in different connectivity modes [71]. Modern energy grids (more contemporarily termed *smart grids*) represent one instance of cyber-physical systems (which are systems intertwining physical and software machine components) which also more generally are often described via hybrid models [72, 73].

As a last example, use cases for hybrid approaches are also amply found in finance and econometrics: exchange rates or stock returns depending on market states can be modeled in this way [74], allowing, e.g., for systematic analyses of regime-switching volatility dynamics [75] and corresponding risk assessment [76, 77].

The study of mathematical hybrid state-space frameworks has a long history in control theory, statistics and machine learning [78–83]. Correspondingly, a great diversity of different model types exists (see, e.g., [80]). The following chapter focuses on MJP-SSDEs as described in Section 2.1.2.3. For MJPs as well as SDEs, the problem of inference has been treated extensively. Exact expressions for the posterior paths given some set of observations can be obtained in each process class, see the discussion in Section 2.2.1.2: classic results include the well-known Kalman and Wonham filters [30], as well as the respective smoothing extensions such as the RTS smoother [21, 55]. However, for diffusion processes in particular, these expressions quickly become intractable as they entail solving multi-dimensional PDEs. This is aggravated if the diffusion is coupled to an underlying jump process, because both processes then have to be solved jointly. Consequentially, a rich variety of approximation methods exists: Monte Carlo approaches (cf. Section 2.2) have been devised for both SDEs, for instance based on particle smoothing [49, 84], and MJPs, e.g., via direct approximations to the jump process [85] or, similarly, particle methods [40, 41, 86]. Similarly, variational methods have been employed both for inference in diffusion processes (utilizing, e.g., Gaussian processes (GPs) [87, 88], moment approximations [43] and general exponential family distributions [23]) and in MJPs [42, 89]. In addition, approximate inference methods also exist for discrete-time SLDS - due to the computational intractability of the full posteriors [11] - such as the Gaussian sum filter [90] or the switching Kalman filter [91]. In the discrete-time domain, SLDS methods currently receive considerable attention [92–94]. Similarly, frameworks utilizing both sampling from the exact posteriors [95, 96] as well as variational approaches [97] have been put forward. In recent years, first inference frameworks for *continuous-time* hybrid systems have also been put forward, see, e.g., [98–100].

In this chapter, a Gibbs sampling scheme is presented which allows to sample from the exact posterior MJP-SSDE process. To this end, the problem is first analyzed mathematically, yielding an evolution equation for the prior process. A similar equation is then derived for the posterior

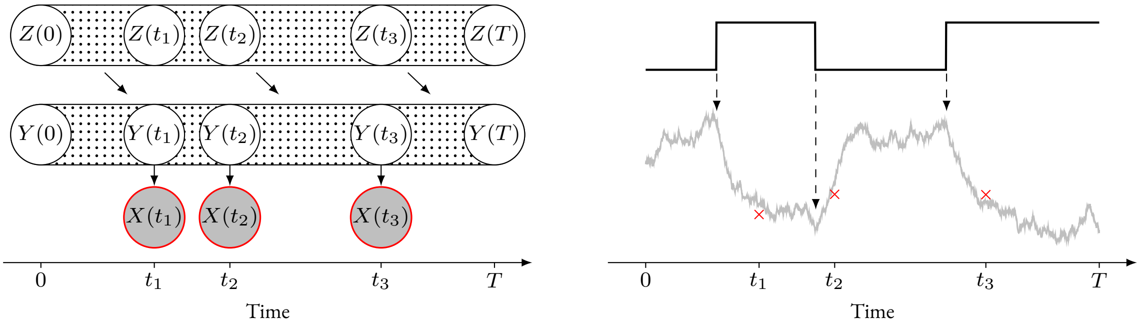


FIGURE 3.1: Sketch of an SSDE hybrid process. Left: graphical model adapted to the continuous-time setting. Right: sketch of a corresponding realization. A switching process Z (top; here, this will be an MJP governed by a transition function Λ throughout), freely evolving in the interval $t \in [0, T]$ controls the dynamics of the SSDE $Y \mid Z$ (middle) via a Z-dependent drift function f and dispersion matrix Q. These latent continuous dynamics generate sparse and noisy observations (bottom, red crosses) at irregularly-spaced time points $t_1, t_2, \ldots$, which are the data available for inference. Vertical dashed arrows indicate the Z-transitions.

process, which is however hard to compute in general. It will be shown that this posterior enables MCMC sampling via a backward-forward/forward-backward-sweeping algorithm. This is combined with a Bayesian treatment of the model parameters, for which full posterior distributions are obtained. The resulting algorithm is first benchmarked on ground-truth data and compared to the exact posterior solution, and then applied to wet-lab fluorescence data on controlled gene-switching. An implementation of the framework is publicly available at `https://git.rwth-aachen.de/bcs/projects/lk/mcmc-ct-sds.git`.

3.1 MARKOV-SWITCHING STOCHASTIC DIFFERENTIAL EQUATIONS

The stochastic dynamical systems considered here consist of MJP-SSDE hybrid processes as defined in Section 2.1.2.3 with discrete observations. In other words, the full model is composed of three joint stochastic processes, cf. Fig. 3.1:

1. a (continuous-time) MJP $Z \coloneqq \{Z(t) : t \in \mathbb{R}_{\geq 0}\}$,

$$Z \sim \mathrm{MJP}(\Lambda(z, z', t)),$$

with $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$,

2. a (continuous-time) subordinated diffusion process, viz., an SSDE, $Y \coloneqq \{Y(t) : t \in \mathbb{R}_{\geq 0}\}$,

$$\mathrm{d}Y(t) = f(Y(t), Z(t), t)\,\mathrm{d}t + Q(Y(t), Z(t), t)\,\mathrm{d}W(t),$$

where $Y(t) \in \mathcal{Y} \subseteq \mathbb{R}^n$, and

3. an observation process $X \coloneqq \{X_i : i \in \mathbb{N}\}$ at discrete time points $\{t_i : i \in \mathbb{N}\}$, $X_i \coloneqq X(t_i) \in \mathcal{X} \subseteq \mathbb{R}^n$.

To avoid ambiguity, in the following the discrete value $Z(t)$ is denoted as the *mode* and the continuous value $Y(t)$ as the *state* of the system.

CHARACTERIZING THE INDUCED DISTRIBUTION For any time interval $[0, T] \subset \mathbb{R}_{\geq 0}$, the MJP-SSDE process induces a measure P on the space Ω_T of all possible paths $\omega_T := \left(y_{[0,T]}, z_{[0,T]}\right)$, where $y_{[0,T]} := \{y(t) : t \in [0, T]\}$, $z_{[0,T]} := \{z(t) : t \in [0, T]\}$ [101]; that is, for any event $\mathcal{A}$ in the Borel σ-algebra of paths, we can formally find its associated probability by integration,

$$\mathsf{P}\left(\left(Y_{[0,T]}, Z_{[0,T]}\right) \in \mathcal{A}\right) = \int_{\mathcal{A}} \mathsf{P}\left(\left(Y_{[0,T]}, Z_{[0,T]}\right) \in \mathrm{d}\omega_T\right) =: \int_{\mathcal{A}} \mathrm{d}\mathsf{P}(\omega_T). \tag{3.1.1}$$

Time point-wise, this quantity admits a probability density $p(y, z, t)$:

$$\mathsf{E}\left[\varphi(Y(t), Z(t), t)\right] = \int_{\Omega} \varphi(y(t), z(t), t)\mathrm{d}\mathsf{P}(\omega_T) = \sum_{z\in\mathcal{Z}} \int_{\mathcal{Y}} \varphi(y, z, t)p(y, z, t)\,\mathrm{d}y, \tag{3.1.2}$$

where $\varphi : \mathcal{Y} \times \mathcal{Z} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is an arbitrary test function. Starting with the Chapman-Kolmogorov equation (2.1.11)

$$p(y, z, t + h) = \sum_{z'} \int_{\mathcal{Y}} p(y, z, t + h \mid y', z', t)p(y', z', t)\,\mathrm{d}y',$$

an evolution equation for this density can be derived by analyzing the properties of the transition density $p(y, z, t + h \mid y', z', t)$ and taking the limit $h \to 0$ as

$$\partial_t p(y, z, t) = \mathcal{L}_t^{\dagger} p(y, z, t), \tag{3.1.3}$$

with some initial distribution $p(y, z, 0)$ and $\mathcal{L}^{\dagger}$ the adjoint of the generator of the hybrid process, cf. Section 2.1.2. Equation (3.1.3) is called the hybrid master equation (HME). It holds that $\mathcal{L}_t^{\dagger} = \mathcal{T}_t^{\dagger} + \mathcal{F}_t^{\dagger}$, where

$$\begin{aligned}
\mathcal{T}_t^{\dagger}\varphi(y, z, t) &:= \sum_{z'\in\mathcal{Z}} \Lambda(z', z, t)\varphi(y, z', t), \\
\mathcal{F}_t^{\dagger}\varphi(y, z, t) &:= -\sum_{i=1}^{n} \partial_{y_i}\left\{f_i(y, z, t)\varphi(y, z, t)\right\} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j}\{D_{ij}(y, z, t)\varphi(y, z, t)\},
\end{aligned}$$

with φ as above and $D(y, z, t) := Q(y, z, t)Q^{\top}(y, z, t)$. The full derivation is provided in Appendix B.1.1, but see also [102].

The same quantity can be obtained as the solution to a PDE going backwards in time, starting at an end point condition $p(y, z, T)$: this backward HME reads

$$\partial_t p(y, z, t) = -\mathcal{L}_t p(y, z, t) \tag{3.1.4}$$

where $\mathcal{L}_t = \mathcal{T}_t + \mathcal{F}_t$ consists of (with again φ as above)

$$\begin{aligned}
\mathcal{T}_t\varphi(y, z, t) &= \sum_{z'\in\mathcal{Z}} \Lambda(z, z', t)\varphi(y, z', t), \\
\mathcal{F}_t\varphi(y, z, t) &= \sum_{i=1}^{n} f_i(y, z, t)\partial_{y_i}\varphi(y, z, t) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(y, z, t)\partial_{y_i}\partial_{y_j}\varphi(y, z, t).
\end{aligned}$$

Inspection of the HME shows that it generalizes - as one would expect - both the density evolution equations for MJPs as well as for SDEs: as Z is independent of Y, marginalizing Eq. (3.1.3) over Y recovers the master equation for MJPs, Eq. (2.1.15),

$$\frac{\mathrm{d}}{\mathrm{d}t} p(z,t) = \sum_{z' \in \mathcal{Z}} \Lambda(z', z, t) p(z,t), \tag{3.1.5}$$

with some $p(z,0)$. Likewise, if $|\mathcal{Z}| = 1$, i.e. in the absence of a switching process, the HME reduces to the conventional FPE,

$$\partial_t p(y,t) = -\sum_{i=1}^{n} \partial_{y_i} \{f_i(y,t) p(y,t)\} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_{y_i} \partial_{y_j} \{D_{ij}(y,z,t) p(y,t)\} \tag{3.1.6}$$

with an initial density $p(y,0)$.

A general, analytic solution to the HME does not exist. One valid approach is to solve this PDE numerically via methods such as finite elements [103]. Solving PDEs is a delicate venture, however, requiring to adapt solvers to the problem at hand and dealing with issues such as the curse of dimensionality and step-size adaptation. On the other hand, sampling trajectories from $\{Y, Z\}$ is straightforward, (cf. Section 2.1.2): a realization $z_{[0,T]}$ of the discrete process Z can be simulated via the Doob-Gillespie algorithm. Given this trajectory, the diffusion Y can be simulated as conventional SDEs using, e.g., an Euler-Maruyama or stochastic Runge-Kutta method.

MODELING ASSUMPTIONS Up to this point, the presented results are general to any MJP-SSDE process. For the remainder of this chapter, the prior MJP is assumed to be time-homogeneous with rate function $\Lambda(z, z', t) = \Lambda(z, z')$. The diffusion components are taken to be mode-dependent, linear and time-invariant, i.e.,

$$f(y,z,t) = f(y,z) = A(z)y + b(z), \tag{3.1.7}$$

with $A(z) \in \mathbb{R}^{n \times n}$ and $b(z) \in \mathbb{R}^n$. Defining the shorthands $\Gamma(z) := [A(z), b(z)] \in \mathbb{R}^{n \times (n+1)}$ and $\bar{y} := \left[y^\top, 1_n^\top\right]^\top \in \mathbb{R}^{n+1}$, where 1_n is the n-dimensional all-ones vector, this may be written as

$$f(y,z) = \Gamma(z)\bar{y}. \tag{3.1.8}$$

Furthermore, the dispersion is assumed to be time-homogeneous and state-independent, i.e., $Q(y,z,t) = Q(z)$. Note that extensions to the time-dependent [24] and state-dependent [23] cases are possible, see also Eq. (2.1.26).

As to the initial distributions, we assume $p(y(0)z(0)) = p(y(0))p(z(0))$ and impose a categorical prior for Z, $p(z(0)) = \mathrm{Cat}(z(0) \mid \pi_0)$, with π_0 a vector of individual entries $\pi_0^i \in [0,1]$, $\sum_i \pi_0^i = 1$. For Y, we assume a Gaussian initial distribution, which we parameterize as $p(y(0)) = \mathcal{N}(y(0) \mid \mu_0, \Sigma_0)$.

Finally, we assume a linear observation model for the data as

$$\begin{aligned} X_i = X(t_i) &= Y(t_i) + \zeta, \\ \zeta &\sim \mathcal{N}(0, \Sigma_x) \end{aligned} \tag{3.1.9}$$

with observation covariance $\Sigma_x \in \mathbb{R}^{n \times n}$.

3.2 EXACT INFERENCE

Consider now a set $x_{[0,T]} := \{x(t_i) : i = 1, \ldots, N\}$ of observations generated via Eq. (3.1.9) at time points $0 \leq t_1, ..., t_N \leq T$. The full inference problem consists in finding the path-wise posterior distribution over the latent hybrid process $\{Y, Z\}$ on the interval $[0, T]$ and all of its parameters Θ given the observed data,

$$\mathsf{P}_x \left((Y_{[0,T]}, Z_{[0,T]}, \Theta) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]}, \theta) \mid x_{[0,T]} \right). \tag{3.2.10}$$

We utilize the subscript x to indicate the posterior measure conditioned on the data $x_{[0,T]}$.

In the smoothing problem, we are (as laid out in Section 2.2.1.1) only interested in the best posterior estimate at a given time point t,

$$\int_\Omega \varphi(Y(t), Z(t), t) \mathrm{d}\mathsf{P}_x = \sum_{z \in \mathcal{Z}} \int_{\mathcal{Y}} \varphi(y, z, t) p(y, z, t \mid x_{[0,T]}) \mathrm{d}y,$$

reducing the problem to computing the smoothing density $p(y, z, t \mid x_{[0,T]})$. By appropriate use of the laws of probability and the model's Markov property, this density can be expressed as the product of two factors with an inherent time-direction: let $k = \max\{k' \in \mathbb{N} : t_{k'} \leq t\}$ for some $t \in [0, T]$, and decompose

$$\begin{aligned} p(y, z, t \mid x_{[0,T]}) &= p(y, z, t, x_{[0,t_k]}) \frac{1}{p(x_{[0,T]})} p(x_{[t_{k+1},T]} \mid y, z, t, x_{[0,t_k]}) \\ &= \frac{p(x_{[0,t_k]})}{p(x_{[0,T]})} p(y, z, t \mid x_{[0,t_k]}) p(x_{[t_{k+1},T]} \mid y, z, t, x_{[0,t_k]}) \\ &= C^{-1}(t) \alpha(y, z, t) \beta(y, z, t). \end{aligned} \tag{3.2.11}$$

The individual components of this expression are the filtering density $\alpha(y, z, t)$, the backward function $\beta(y, z, t)$, and the normalizer $C(t)$:

$$\begin{aligned} \alpha(y, z, t) &= p(y, z, t \mid x_{[0,t_k]}), \\ \beta(y, z, t) &= p(x_{[t_{k+1},T]} \mid y, z, t), \\ C(t) &= \sum_z \int \alpha(y, z, t) \beta(y, z, t) \, \mathrm{d}y. \end{aligned} \tag{3.2.12}$$

The (forward) filtering density is indeed a proper probability density, which is specificed at some - with respect to the given data - "future" time point $t \geq t_k$. On the other hand, the backward function - as the name suggests - is not a density with respect to y, z. Relative to the given data $x_{[t_{k+1},T]}$, it provides information about a "past" time point $t \leq t_{k+1}$. These functions are similar to the ones obtained for the continuous-discrete HMM, see Section 2.2.1. In the following, it will be shown how the exact inference problem can be solved via Eq. (3.2.11).

First, notice that the normalizer $C(t)$ is constant with probability 1 [23, 104]; between observations, it is given via the fixed ratio $C(t)^{-1} = p(x_{[0,t_k]})/p(x_{[0,T]})$ for $t_k \leq t < t_{k+1}$. The dynamics of Eq. (3.2.11) are thus prescribed by the dynamics of the forward and backward densities:

$$\partial_t p(y, z, t \mid x_{[0,T]}) = C^{-1}(t) \big(\alpha(y, z, t) \partial_t \beta(y, z, t) + \beta(y, z, t) \partial_t \alpha(y, z, t) \big). \tag{3.2.13}$$

It is only at the observation times that the normalizer changes instantaneously upon "incorporation" of the next observed datum. As is typical for continuous-discrete problems [20], these quantities hence (i) follow the prior dynamics *between* and (ii) are subject to jump conditions *at* the observation time points.

DYNAMICS BETWEEN OBSERVATIONS The smoothing dynamics between observation time points are straightforwardly derived via the system's Markov property. Consider a time interval $[t, t+h], h > 0$ in which *no* observation occurs. With the index k defined as above (cf. Eq. (3.2.11)), we have by the law of total probability

$$\alpha(y, z, t+h) = \sum_{z' \in \mathcal{Z}} \int p(y, z, t+h \mid y', z', t, x_{[0,t_k]}) p(y', z', t \mid x_{[0,t_k]}) \, \mathrm{d}y',$$

and the Markov property ensures

$$p(y, z, t+h \mid y', z', t, x_{[0,t_k]}) = p(y, z, t+h \mid y', z', t).$$

Thus,

$$\alpha(y, z, t+h) = \sum_{z' \in \mathcal{Z}} \int p(y, z, t+h \mid y', z', t) \alpha(y', z', t) \, \mathrm{d}y',$$

which is the usual Chapman-Kolmogorov equation for the filtering process α with transition distribution $p(y, z, t+h \mid y', z', t)$. As this is the transition distribution of the prior dynamics (3.1.3), the filtering distribution follows the same HME between observations,

$$\partial_t \alpha(y, z, t) = \mathcal{L}_t^\dagger \alpha(y, z, t), \tag{3.2.14}$$

with some initial condition $\alpha(y, z, 0)$.

Similarly, for the backward function $\beta(y, z, t)$, consider an interval without observations $[t - h, t], h > 0$. In analogy to the above,

$$\beta(y, z, t-h) = \sum_{z' \in \mathcal{Z}} \int p(y', z', t \mid y, z, t-h) p(x_{[t_{k+1}, T]} \mid y', z', t, y, z, t-h) \, \mathrm{d}y'$$

and

$$p(x_{[t_{k+1}, T]} \mid y', z', t, y, z, t-h) = p(x_{[t_{k+1}, T]} \mid y', z', t),$$

which yields

$$\beta(y, z, t-h) = \sum_{z' \in \mathcal{Z}} \int p(y', z', t \mid y, z, t-h) \beta(y', z', t) \, \mathrm{d}y'.$$

This is the Chapman-Kolmogorov equation for $\{x_{k+1}, \ldots, x_N \mid Y(t), Z(t), t \leq t_{k+1}\}$ - i.e. the backward process - with the transition distribution $p(y', z', t \mid y, z, t-h)$ corresponding to the backward prior dynamics. Accordingly, the backward function $\beta(y, z, t)$ follows the backward HME

$$\partial_t \beta(y, z, t) = -\mathcal{L}_t \beta(y, z, t), \tag{3.2.15}$$

in between observations with the end point condition $\beta(y, z, T) = 1$.

JUMP CONDITIONS The behavior at an observation time point t_k is found by separating the contributions of the filtering distribution conditioned on all *previous* observations and the observation likelihood of $x_k = x(t_k)$:

$$\begin{aligned}
\alpha(y,z,t_k) &= p(y,z,t_k \mid x_{[0,t_k]}) \\
&= \frac{p(y,z,t_k,x_{[0,t_k]})}{p(x_{[0,t_k]})} \\
&= \frac{p(x_k \mid y,z,t_k,x_{[0,t_{k-1}]})p(y,z,t_k,x_{[0,t_{k-1}]})}{p(x_{[0,t_k]})} \\
&= \frac{p(x_k \mid y,z,t_k,x_{[0,t_{k-1}]})p(y,z,t_k \mid x_{[0,t_{k-1}]})p(x_{[0,t_{k-1}]})}{p(x_{[0,t_k]})} \\
&= \frac{p(x_k \mid y,t_k)\alpha(y,z,t_k^-)}{C_k}
\end{aligned} \tag{3.2.16}$$

with $C_k := \frac{p(x_{[0,t_k]})}{p(x_{[0,t_{k-1}]})} = \sum_{z\in\mathcal{Z}} \int p(x_k \mid y)\alpha(y,z,t_k^-)\,\mathrm{d}y$ and

$$\alpha(y,z,t_k^-) = \lim_{h\searrow 0} \alpha(y,z,t_k-h), \tag{3.2.17}$$

the value of the filter *right before the jump*, that is, the value of the solution of Eq. (3.2.14) at t_k.

Analogously, one computes

$$\begin{aligned}
\beta(y,z,t_k) &= p(x_{[t_k,T]} \mid y,z,t_k) \\
&= \frac{p(x_{[t_k,T]},y,z,t_k)}{p(y,z,t_k)} \\
&= \frac{p(x_k \mid x_{[t_{k+1},T]},y,z,t_k)p(x_{[t_{k+1},T]},y,z,t_k)}{p(y,z,t_k)} \\
&= p(x_k \mid y,t_k)\beta(y,z,t_k^+)
\end{aligned} \tag{3.2.18}$$

where

$$\beta(y,z,t_k^+) = \lim_{h\searrow 0} \beta(y,z,t_k+h). \tag{3.2.19}$$

The PDEs (3.2.14) and (3.2.15), together with the respective jump conditions (3.2.16) and (3.2.18), constitute a set of *impulsive* differential equations [105–107]. To solve these equations, one proceeds as the jump conditions suggest: first, given some initial (or terminal) conditions, solve Eq. (3.2.14) (or Eq. (3.2.15)) until the next observation time point; second, apply the jump conditions Eq. (3.2.16) (or Eq. (3.2.18)), and third, repeat this process - with the value after applying the jump condition as the new initial (terminal) value - until all observations are incorporated.

By inserting the dynamics of $\alpha(y,z,t)$ and $\beta(y,z,t)$ into Eq. (3.2.11), it can readily be shown that the smoothing distribution itself follows a HME. The calculations are provided in Appendix B.1.2, resulting in

$$\partial_t p(y,z,t \mid x_{[0,T]}) = \tilde{\mathcal{L}}_t^\dagger p(y,z,t \mid x_{[0,T]}), \tag{3.2.20}$$

with initial condition $p(y, z, 0 \mid x_{[0,T]}) \propto \alpha(y, z, 0)\beta(y, z, 0)$ and $\tilde{\mathcal{L}}_t^\dagger = \tilde{\mathcal{T}}_t^\dagger + \tilde{\mathcal{F}}_t^\dagger$,

$$\begin{aligned}
\tilde{\mathcal{T}}_t^\dagger \varphi(y, z, t) &:= \sum_{z' \in \mathcal{Z}} \tilde{\Lambda}(y, z', z, t)\varphi(y, z', t) \\
\tilde{\mathcal{F}}_t^\dagger \varphi(y, z, t) &:= -\sum_{i=1}^{n} \partial_{y_i} \left\{ \tilde{f}_i(y, z, t)\varphi(y, z, t) \right\} \\
&\qquad\qquad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j}\{\tilde{D}_{ij}(y, z, t)\varphi(y, z, t)\},
\end{aligned}$$

where the posterior drift, dispersion and rate function

$$\tilde{f}_i(y, z, t) = f_i(y, z, t) + \sum_{j=1}^{n} D_{ij}(y, z, t)\partial_{y_j}\{\ln \beta(y, z, t)\}, \tag{3.2.21}$$

$$\tilde{D}(y, z, t) = D(y, z, t), \tag{3.2.22}$$

$$\tilde{\Lambda}(y, z', z, t) = \Lambda(z', z, t)\frac{\beta(y, z, t)}{\beta(y, z', t)}, \tag{3.2.23}$$

and φ an arbitrary test function as above. The posterior density can hence in principle be evaluated by first solving the backward dynamics of β, Eq. (3.2.15) starting at $t = T$, and subsequently computing the smoothing HME (3.2.20) from $t = 0$.

3.3 INFERENCE VIA MARKOV CHAIN MONTE CARLO

As the exact smoothing problem accordingly requires the solution of two coupled PDEs with discrete and continuous components, its computation can be challenging. Yet, their solution only provides the time point-wise posterior density. In this section, a blocked Gibbs sampler is presented that avoids this need to compute two PDEs and yields full path-space information, allowing one, e.g., to compute correlations over time. The switching process $Z_{[0,T]}$, the diffusion process $Y_{[0,T]}$ and the parameters Θ are sampled in turn from the complete conditional measures

$$Y_{[0,T]} \sim \mathsf{P}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]}, x_{[0,T]}, \theta), \tag{3.3.24}$$

$$Z_{[0,T]} \sim \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]}, x_{[0,T]}, \theta), \tag{3.3.25}$$

$$\Theta \sim \mathsf{P}(\Theta \in \mathrm{d}\theta \mid y_{[0,T]}, z_{[0,T]}, x_{[0,T]}). \tag{3.3.26}$$

Hence, this scheme yields samples from the full posterior path measure (3.2.10), cf. Section 2.2.4.1.

The measures (3.3.24) and (3.3.25) can be shown to each describe conditional Markov processes. Drawing on results from filtering and smoothing theory, we now derive respective evolution equations on the process level, allowing for the generation of the desired samples. By using conjugate prior distributions, Eq. (3.3.26) yields closed-form distributions for all parameters; notably, for the dispersion Q, we do not obtain the posterior directly, but utilize a Metropolis-adapted Langevin scheme, ensuring numerical stability [108]. The conditional measures are derived in order in the following.

3.3.1 *Sampling the conditional diffusion process*

To generate samples from the full conditional diffusion measure

$$Y_{[0,T]} \sim \mathsf{P}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]}, x_{[0,T]}, \theta),$$

first acknowledge that by conditioning on the trajectory $z_{[0,T]}$, the SSDE model can be interpreted as a temporal sequence of conventional SDEs, or, equivalently - as all drift functions Eq. (3.1.7) are assumed to be linear - as one SDE with time-dependent parameters. More concretely, we re-interpret

$$\begin{aligned} f : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^n,& \\ f(y,z) = A(z)y + b(z),& \end{aligned}$$

as

$$\begin{aligned} f : \mathcal{Y} \times [0,T] \to \mathbb{R}^n,& \\ f(y,t) = A(z(t))y(t) + b(z(t)),& \end{aligned}$$

and analogously for the dispersion.

The path measure Eq. (3.3.24) can hence be obtained as the posterior of an SDE process. For a conventional (i.e., non-switching) SDE

$$\mathrm{d}Y(t) = f(Y(t),t)\mathrm{d}t + Q(t)\mathrm{d}W(t),$$

it is well-known that the posterior process conditioned on some data $x_{[0,T]}$ can in turn be expressed as an SDE [106],

$$\mathrm{d}Y(t) = \tilde{f}(Y(t),t)\mathrm{d}t + Q(t)\mathrm{d}W(t), \tag{3.3.27}$$

with the posterior drift function $\tilde{f}(y,t)$:

$$\tilde{f}(y,t) = f(y,t) + D(t)\partial_y \ln \beta(y,t), \tag{3.3.28}$$

where $\beta(y,t) = p(x_{[t_{k+1},T]} \mid y,t)$ and $k = \max\{k' \in \mathbb{N} : t_{k'} \leq t\}$ similar to the above.

Notice how this also follows readily from the posterior HME derived in Section 3.2 by setting $|\mathcal{Z}| = 1$. The backward PDE (3.2.15) in this case reduces to the KBE (see Section 2.1.2) starting at the end-point $t = T$ with $\beta(y,T) = 1$:

$$\partial_t \beta(y,t) = -\sum_{i=1}^{n} f_i(y,t)\partial_{y_i}\beta(y,t) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(t)\partial_{y_i}\partial_{y_j}\beta(y,t). \tag{3.3.29}$$

Under the given modeling assumptions, this KBE can be evaluated in closed form; this yields a Kalman-type backward filter which has also been derived in the context of smoothing for nonlinear diffusions in [109]. This allows us to employ a backward-filtering, forward-sampling approach: first, solve Eq. (3.3.29) backwards starting at $t = T$. Second, given this solution, sample a trajectory of the posterior SDE (3.3.27) forward in time.

The full derivation of the solution of Eq. (3.3.29) is provided in Appendix B.2.1: utilizing the method of characteristics [110] on the Fourier transform of $\beta(y,t)$ we find

$$\beta(y,t) = \mathcal{N}\left(x_{[t_k,T]} \mid F(t)y + m(t), \Sigma(t)\right), \, t < t_k, \tag{3.3.30}$$

where $F(t) \in \mathbb{R}^{(N-k)n \times n}$, $m(t) \in \mathbb{R}^{(N-k)n}$, and $\Sigma(t) \in \mathbb{R}^{(N-k)n \times (N-k)n}$ are determined by a set of ODEs:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F(t) &= -F(t)A(t) && \text{with} && F(T) = F, \\ \frac{\mathrm{d}}{\mathrm{d}t} m(t) &= -F(t)b(t) && \text{with} && m(T) = 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} \Sigma(t) &= -F(t)D(t)F(t)^\top && \text{with} && \Sigma(T) = \Sigma. \end{aligned} \tag{3.3.31}$$

These ODEs are subject to the jump conditions (3.2.18), yielding

$$\begin{aligned} F(t_{k-1}) &= \begin{pmatrix} \mathbb{1}_{n\times n} \\ F(t_{k-1}^+) \end{pmatrix} \in \mathbb{R}^{(N-(k-1))n \times n}, \\ m(t_{N-1}) &= \begin{pmatrix} 0 \\ m(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{(N-(k-1))n}, \\ \Sigma(t_{N-1}) &= \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{(N-(k-1))n \times (N-(k-1))n}, \end{aligned} \tag{3.3.32}$$

where $\mathbb{1}_{n\times n}$ is the n-dimensional identity matrix and $F(t_{k-1}^+)$, $m(t_{N-1}^+)$ and $\Sigma(t_{N-1}^+)$ are defined as Eq. (3.2.19).

The support of this distribution increases with each incorporated observation, which is computationally inconvenient. This issue can be solved by interpreting the Gaussian (3.3.30) as a distribution over y rather than x: by completing the square, we find

$$\ln \beta(y,t) = -c(t) - \frac{1}{2} y^\top I(t) y + a(t)^\top y, \tag{3.3.33}$$

with

$$\begin{aligned} a(t) &:= F(t)^\top \Sigma^{-1}(t)(x(t) - m(t)), \\ I(t) &:= F(t)^\top \Sigma^{-1}(t) F(t), \end{aligned} \tag{3.3.34}$$

and $c(t)$ is a normalizer. As the posterior drift (3.3.27) only depends on the gradient of Eq. (3.3.33),

$$\partial_y \ln \beta(y,t) = -I(t)y + a(t), \tag{3.3.35}$$

this normalizer is irrelevant for sampling from the posterior.

Under this reparameterization, the KBE solution (3.3.31) yields a continuous-time analogue to the discrete-time information filter [111]:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} I(t) &= -A(t)^\top I(t) - I(t)A(t) + I(t)D(t)I(t), \\ \frac{\mathrm{d}}{\mathrm{d}t} a(t) &= -A(t)^\top a(t) + I(t)D(t)a(t) + I(t)b(t). \end{aligned} \tag{3.3.36}$$

Importantly, these parameters are fixed in size: $I(t) \in \mathbb{R}^{n \times n}$ and $a(t) \in \mathbb{R}^n$. The jump conditions follow readily by comparing Eq. (3.3.32) and Eq. (3.3.34) as

$$\begin{aligned} I(t_i) &= \Sigma_x^{-1} + I(t_i^+), \\ a(t_i) &= \Sigma_x^{-1} x_i + a(t_i^+). \end{aligned} \tag{3.3.37}$$

We solve these ODEs by utilizing standard numerical adaptive step-size solvers [25, 112].

Having computed $\partial_y \ln \beta(y, t)$ backwards from $t = T$ to $t = 0$, we can now straightforwardly simulate the SDE (3.3.24) forward in time, yielding samples from the full conditional distribution in Eq. (3.2.10). To initialize the sample trajectory, notice that the posterior of the initial value $y(0)$, $p(y(0) \mid x_{[0,T]})$, is readily found as

$$\begin{aligned} p(y(0) \mid x_{[0,T]}, -) &\propto \beta(y(0), 0) p(y(0) \mid \mu_0, \Sigma_0) \\ &\propto \exp\left\{-\frac{1}{2} y(0)^\top I(0) y(0) + a(0)^\top y(0)\right\} \mathcal{N}(y(0) \mid \mu_0, \Sigma_0) \\ &\propto \mathcal{N}\big(y(0) \mid \bar{\mu}, \bar{\Sigma}\big) \end{aligned} \tag{3.3.38}$$

with

$$\bar{\mu} = \bar{\Sigma}(\Sigma_0^{-1} \mu_0 + a(0)), \quad \bar{\Sigma} = \left(\Sigma_0^{-1} + I(0)\right)^{-1}. \tag{3.3.39}$$

3.3.2 *Sampling the conditional switching process*

Given the simulated SSDE path $y_{[0,T]}$, we want to sample from the full conditional switching path measure,

$$Z_{[0,T]} \sim \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]}, x_{[0,T]}, \theta),$$

which due to the Markovian structure described in Section 3.1 reduces to

$$Z_{[0,T]} \sim \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]}, \theta). \tag{3.3.40}$$

This setting is qualitatively different from the continuous-discrete smoothing problem in the previous section: the observations now consist of a *full path* instead of a finite set of points. Still, as the classic continuous-discrete MJP smoothing problem can be solved in the same way as the just described SSDE-diffusion [32], one would expect that sampling the conditional switching process should analogously be possible by backward-filtering and forward-sampling. It will be shown that this is indeed the case, but the resulting expressions are no longer ODEs with jump conditions.

First, notice that conditioning on some path $y_{[0,T]}$ does not alter the Markovian structure of the model; that is, Eq. (3.3.40) is given via an MJP [113]. To be able to simulate $z_{[0,T]}$, we hence desire to know the dynamics of the smoothing marginal

$$\begin{aligned} p_s(z, t) &:= \mathsf{E}\left[\mathbb{1}(Z(t) = z) \mid y_{[0,T]}\right] \\ &= \int \mathbb{1}(z(t) = z) \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]}). \end{aligned} \tag{3.3.41}$$

As detailed in Section 2.2.1, we need to resort to the Kallianpur-Striebel formula (which is the equivalent of Bayes' rule for continuous stochastic processes) to obtain the posterior measure $\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]})$. First, we know (cf. Section 2.2.1) that the conditional measure $\mathsf{P}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})$ and the measure $\mathsf{P}(W_{[0,T]} \in \mathrm{d}y_{[0,T]})$ of standard Brownian motion relate via the Radon-Nikodym derivative

$$\begin{aligned} G(y_{[0,T]}, z_{[0,T]}) :=& \frac{\mathsf{P}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})}{\mathsf{P}(W_{[0,T]} \in \mathrm{d}y_{[0,T]})} \\ =& \exp\left\{ \int_0^T f(y(s), z(s))^\top D^{-1}(z(s)) \mathrm{d}y(s) \right. \\ & \left. -\frac{1}{2} \int_0^T f(y(s), z(s))^\top D^{-1}(z(s)) f(y(s), z(s)) \mathrm{d}s \right\}. \end{aligned} \tag{3.3.42}$$

This allows to replace the conditional measure with standard Brownian motion as

$$\begin{aligned} \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]}) &\propto \mathsf{P}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})\, \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}) \\ &\propto G(y_{[0,T]}, z_{[0,T]})\, \mathsf{P}(W_{[0,T]} \in \mathrm{d}y_{[0,T]})\, \mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}), \end{aligned}$$

yielding the desired Kallianpur-Striebel equation as

$$\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]} \mid y_{[0,T]}) = \frac{G(y_{[0,T]}, z_{[0,T]})\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]})}{\int G(y_{[0,T]}, z'_{[0,T]})\mathsf{P}(Z_{[0,T]} \in \mathrm{d}z'_{[0,T]})}. \tag{3.3.43}$$

Starting with Eq. (3.3.41), the general idea is - akin to the preceding section - to compute the time point-wise marginal, and then separate the resulting expression into a forward and a backward component for which individual evolution expressions can be derived. To that end, first note that

$$G(y_{[0,T]}, z_{[0,T]}) = G(y_{[0,t]}, z_{[0,t]}) G(y_{[t,T]}, z_{[t,T]}). \tag{3.3.44}$$

With this, Eqs. (3.3.41) and (3.3.43) yield (for the unnormalized version $\tilde{p}_s$ of the smoothing density)

$$\begin{aligned} &\tilde{p}_s(z,t) \\ &\quad = \mathsf{E}\big[\mathbb{1}(Z(t)=z) G(Y_{[0,t]}, Z_{[0,t]}) G(Y_{[t,T]}, Z_{[t,T]}) \bigm| y_{[0,T]}\big] \\ &\quad = \mathsf{E}\big[\mathbb{1}(Z(t)=z) G(Y_{[0,t]}, Z_{[0,t]})\, \mathsf{E}[G(Y_{[t,T]}, Z_{[t,T]}) \mid Z(t), Y_{[t,T]} = y_{[t,T]}] \bigm| y_{[0,T]}\big] \\ &\quad = \mathsf{E}\left[\mathbb{1}(Z(t)=z) G(Y_{[0,t]}, Z_{[0,t]}) \mid y_{[0,t]}\right] \mathsf{E}[G(Y_{[t,T]}, Z_{[t,T]}) \mid y_{[t,T]}, Z(t) = z], \end{aligned} \tag{3.3.45}$$

where the expectations are taken with respect to the prior measure and $y_{[0,T]}$ is the given, simulated diffusion path. Inspecting both expectations individually, we find

$$\begin{aligned} &\mathsf{E}\left[\mathbb{1}(Z(t)=z) G(Y_{[0,t]}, Z_{[0,t]}) \mid y_{[0,t]}\right] \\ &\qquad = \int \mathbb{1}(z(t)=z) G(y_{[0,t]}, z_{[0,t]}) \mathsf{P}(Z_{[0,t]} \in \mathrm{d}z_{[0,t]}) \\ &\qquad \propto \int \mathbb{1}(z(t)=z) \mathsf{P}(Z_{[0,t]} \in \mathrm{d}z_{[0,t]} \mid y_{[0,t]}) =: \tilde{p}_f(z,t), \end{aligned} \tag{3.3.46}$$

the (accordingly unnormalized) filtering density, and we notice that

$$\mathsf{E}[G(Y_{[t,T]}, Z_{[t,T]}) \mid y_{[t,T]}, Z(t) = z] =: v(z,t), \tag{3.3.47}$$

is the continuous analogue to the backward function β Eq. (3.2.12): we have

$$\tilde{p}_s(z,t) = \tilde{p}_f(z,t) v(z,t).$$

To obtain the dynamics of $\tilde{p}_s$ (and ultimately p_s), we need to compute the derivative of this expression. The stochastic integral obeys a generalized version of the conventional integration by parts [13],

$$\mathrm{d}\tilde{p}_s(z,t) = v(z,t)\mathrm{d}\tilde{p}_f(z,t) + \tilde{p}_f(z,t)\mathrm{d}v(z,t) + \mathrm{d}\tilde{p}_f(z,t)\mathrm{d}v(z,t). \tag{3.3.48}$$

Recall that, as discussed in Chapter 2, Itô's lemma also applies for jump processes; hence, we can proceed to express the dynamics of $\tilde{p}_s(z,t)$ via the individual dynamics of $\tilde{p}_f(z,t)$ and $v(z,t)$. Both will be stated in the following without a detailed derivation, which can however be found in Appendix B.3.

BACKWARD DYNAMICS Via Itô calculus, we find

$$\begin{aligned}\mathrm{d}v(z,t) = -v(z,t) f(y,z,t)^\top D^{-1}(t)\mathrm{d}y(t) - \sum_{z'} v(z',t)\Lambda(z,z',t)\mathrm{d}t \\ + f(y,z,t)^\top D^{-1}(t) f(y,z,t) v(z,t)\mathrm{d}t.\end{aligned} \tag{3.3.49}$$

Notice that we write $\mathrm{d}y(t)$ instead of $\mathrm{d}Y(t)$, as we are conditioning on the previously drawn sample path $y_{[0,T]}$.

FORWARD DYNAMICS Similarly, we obtain for the unnormalized filtering dynamics the *Zakai equation*

$$\mathrm{d}\tilde{p}_f(z,t) = \sum_{z' \in \mathcal{Z}} \Lambda(z',z,t)\tilde{p}_f(z',t)\mathrm{d}t + \tilde{p}_f(z,t) f(y(t),z(t))^\top D^{-1}(z,t)\mathrm{d}y(t). \tag{3.3.50}$$

The corresponding normalized filtering dynamics are obtained from the Zakai equation as the *Kushner-Stratonovich equation* [13]

$$\begin{aligned}\mathrm{d}p_f(z,t) = \sum_{z' \in \mathcal{Z}} \Lambda(z',z,t) p_f(z',t)\mathrm{d}t \\ + p_f(z,t)(f(y,z) - \bar{f}(y,t))^\top D^{-1}(z,t)(\mathrm{d}y(t) - \bar{f}(y,t)\mathrm{d}t)\end{aligned} \tag{3.3.51}$$

with $\bar{f}(y,t) = \sum_{z' \in \mathcal{Z}} f(z',y) p_f(z',t)$. Note that for $f(y,z) = f(z)$, this recovers the classic Wonham filter [114].

Inserting both above results into Eq. (3.3.45), one obtains

$$\begin{aligned}\mathrm{d}\tilde{p}_s(z,t) &= v(z,t)\sum_{z'} \Lambda_{z'z}(t)\tilde{p}_f(z',t)\mathrm{d}t - \sum_{z'} v(z',t)\Lambda_{zz'}(t)\tilde{p}_f(z,t)\mathrm{d}t \\ &= \frac{\tilde{p}_s(z,t)}{\tilde{p}_f(z,t)}\sum_{z'} \Lambda_{z'z}(t)\tilde{p}_f(z',t)\mathrm{d}t - \sum_{z'} \frac{\tilde{p}_s(z',t)}{\tilde{p}_f(z',t)}\Lambda_{zz'}(t)\tilde{p}_f(z,t)\mathrm{d}t,\end{aligned} \tag{3.3.52}$$

where Eq. (3.3.45) is used to replace $v(z,t)$. As the filtering densities $\tilde{p}_f$ only occur as ratios, it is permissible to readily replace $\tilde{p}_f \to p_f$. Summation over z shows

$$\sum_z \mathrm{d}\tilde{p}_s(z,t) = 0,$$

implying that p_s has a time-independent normalizer. Consequently,

$$\mathrm{d}p_s(z,t) = \frac{p_s(z,t)}{p_f(z,t)} \sum_{z'} \Lambda_{z'z}(t) p_f(z',t)\mathrm{d}t - \sum_{z'} \frac{p_s(z',t)}{p_f(z',t)} \Lambda_{zz'}(t) p_f(z,t)\mathrm{d}t. \tag{3.3.53}$$

It can be seen by defining the *posterior rate function*

$$\begin{aligned} \tilde{\Lambda}(z',z,t) &= \frac{p_f(z,t)}{p_f(z',t)} \Lambda(z,z'), z' \neq z, \\ \tilde{\Lambda}(z,z,t) &:= -\sum_{z' \neq z} \tilde{\Lambda}(z,z',t), \end{aligned} \tag{3.3.54}$$

that this has the form of a backward master equation (cf. Section 2.1.2.1):

$$\frac{\mathrm{d}}{\mathrm{d}t} p_s(z,t) = -\sum_{z' \in \mathcal{Z}} \tilde{\Lambda}(z',z,t) p_s(z',t). \tag{3.3.55}$$

This finally allows to backward-sample a new path $z_{[0,T]}$ with the end-point condition $p_s(z,T) = p_f(z,T)$ after forward-filtering via Eq. (3.3.51). This temporal order is computationally advantageous: it allows us to simultaneously solve all occurring stochastic integrals of the joint problem $\{Y, Z \mid x_{[0,T]}\}$ in the *same time direction*. In other words, the path $y_{[0,T]}$ and the filtering distribution $\tilde{p}_f$ can be computed simultaneously - this saves one full pass through the trajectory compared to an approach where $y_{[0,T]}$ and $z_{[0,T]}$ would be computed successively by a backward-forward scheme; see also the illustration in Fig. 3.2. To simulate the conditional switching process Z with time-dependent rates Eq. (3.3.54), the thinning algorithm is utilized, cf. Section 2.1.2.1.

Note that it is possible in principle to reverse this procedure and backward-filter/forward-sample as for the diffusion component $y_{[0,T]}$. This however includes defining a *reverse* diffusion process $\overleftarrow{Y}(t)$ which evolves from $t = T$ to $t = 0$ and accounts for the asymmetry of the Itô integral; for more details on this, the interested reader is referred to, e.g. [31, 115].

The above formulation of the backward and forward dynamics consists of expressions based on stochastic integrals with respect to $Y_{[0,T]}$. In practice, these can only be solved approximately (e.g. via the Euler-Maruyama scheme). Mathematically, this raises questions, for example with respect to the continuity of solutions under variations of the model parameters. These questions can in principle be more appropriately dealt with via so-called *robust* solutions that are not expressed as stochastic integrals, but as proper functions of the path realizations $y_{[0,T]}$. Appropriate theoretical results exist for pure SDE systems with state-independent drift [55, 104]; see in particular the latter reference for a thorough discussion. It is, however, unclear whether such a robust solution can exist at all for hybrid models. To provide some intuition about this issue, a short digression follows that may point the interested reader to potential topics of further inquiry. Readers wanting to continue with the next (and last) sampling step concerning the system parameters Θ may want to fast-forward to p. 41.

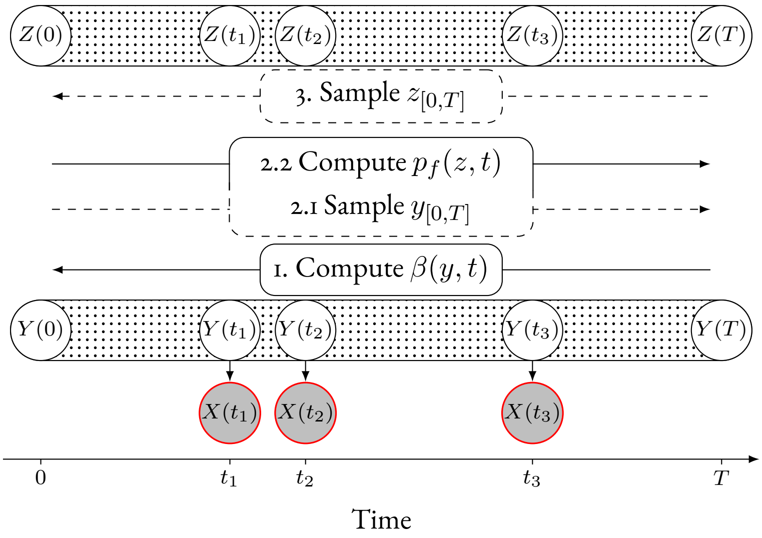


FIGURE 3.2: Sketch of the backward-forward/forward-backward sampling scheme. First, the backward function $\beta(y,t)$ is computed from $t = T$ to $t = 0$ with the information filter Eq. (3.3.36). Second, the new posterior path $y_{[0,T]}$ is sampled forward in time via Eq. (3.3.27); this can be done simultaneously with the computation of the forward filtering density $p_f(z,t)$ (or, equivalently, its unnormalized counterpart) via the Kushner-Stratonovich Eq. (3.3.51) (or the Zakai equation (3.3.50)). Lastly, the new posterior path $z_{[0,T]}$ can be sampled backwards from $t = T$ to $t = 0$ via Eq. (3.3.55).

DIGRESSION: ROBUST FILTERING To gain some intuition about this approach as well as the problems arising in the present context, consider the conventional 1D filtering setting as described in [116] (but see also [13, 31, 55]) with a latent jump process and a dependent observed diffusion process,

$$\begin{aligned} Z &\sim \mathrm{MJP}(\Lambda(z,z')), \\ \mathrm{d}Y(t) &= h(Z(t))\mathrm{d}t + \mathrm{d}W(t), \end{aligned}$$

with $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$, $h : \mathcal{Z} \to \mathbb{R}$ and W a one-dimensional Brownian motion. The unnormalized filtering density is described as above by the Zakai equation, which reads

$$\mathrm{d}\tilde{p}_f(z,t) = \sum_{z' \in \mathcal{Z}} \Lambda(z',z)\tilde{p}_f(z',t)\mathrm{d}t + h(z)\tilde{p}_f(z,t)\mathrm{d}Y(t).$$

Converting the Itô-type to a Stratonovich-type stochastic integral [20], we have

$$\mathrm{d}\tilde{p}_f(z,t) = \left(\sum_{z' \in \mathcal{Z}} \Lambda(z',z) + \frac{1}{2}h(z)^2\right)\tilde{p}_f(z',t) + h(z)\tilde{p}_f(z,t) \circ \mathrm{d}Y(t), \tag{3.3.56}$$

for which (in contrast to Itô integrals) the differentiation rules of conventional calculus apply. It can thereby be easily verified that the solution to this equation given some observation path $y_{[0,T]}$ is found via the ansatz

$$\tilde{p}_f(z,t) = r(t)\exp\{h(z)y(t)\}, \tag{3.3.57}$$

resulting in Eq. (3.3.56) being fulfilled if

$$\frac{\mathrm{d}}{\mathrm{d}t}r(t) = \exp\{-h(z)y(t)\}\left(\Lambda(z',z) - \frac{1}{2}h(z)^2\right)\exp\{h(z')y(t)\}\,r(z',t).$$

This is a "robust" solution because it is given as an ODE parameterized by the full observed path; specifically, it is not given as an expectation hinging on stochastic integrals. If however $h = h(z, y)$, this does not work any more: computing the time derivative of Eq. (3.3.57) shows that additional terms in $\frac{\mathrm{d}}{\mathrm{d}t} y(t)$ emerge that do not cancel out in Eq. (3.3.56) and hence do not yield a resulting ODE with a smooth dependence on the full path. This issue falls into the realm of *rough path theory*, which finds that already for multivariate, but state-independent observation functions $h(z)$, robust formulations may not exist [116, 117]. For the present, even more general case, the additional state-dependence of the SSDE drift function hence does not seem to admit a robust ODE formulation of the backward function v.

3.3.3 *Sampling the parameters*

The presented inference framework is naturally complemented with Bayesian parameter estimation. To this end, conjugate prior distributions are specified over the model parameters. In the following, the resulting full conditionals are derived.

INITIAL DISTRIBUTIONS On the initial MJP state distribution parameter π_0, a Dirichlet prior with hyperparameter $\alpha_{\pi_0} \in \mathbb{R}_{>0}^{|\mathcal{Z}|}$ is imposed,

$$p(\pi_0) = \mathrm{Dir}(\pi_0 \mid \alpha_{\pi_0}), \tag{3.3.58}$$

yielding

$$\begin{aligned} p(\pi_0 \mid z_{[0,T]}, -) &= \mathrm{Cat}(z(0) \mid \pi_0)\,\mathrm{Dir}(\pi_0 \mid \alpha_{\pi_0}) \\ &= \mathrm{Dir}(\pi_0 \mid \alpha_{\pi_0} + \delta_{z(0)}) \end{aligned} \tag{3.3.59}$$

with the point mass on $z(0)$, $\delta_{z(0)}$. Note that all variables in the conditioning set that π_0 is conditionally independent of are suppressed. This convention is followed in all update equations for conciseness.

On the SSDE initial distribution parameters μ_0, Σ_0, a Normal-inverse-Wishart (NIW) prior is placed:

$$\mu_0, \Sigma_0 \sim \mathrm{NIW}(\eta, \rho, \Psi, \nu) \Leftrightarrow \begin{cases} \mu_0 \sim \mathcal{N}(\eta, \Sigma_0/\rho), \\ \Sigma_0 \sim \mathrm{IW}(\Psi, \nu), \end{cases} \tag{3.3.60}$$

where the inverse-Wishart (IW) distribution is characterized by the density function

$$\mathrm{IW}(\Sigma \mid \Psi, \nu) = \frac{|\Psi|^{\nu/2} |\Sigma|^{-(\nu+n+1)/2}}{2^{n\nu/2} \Gamma_n(\frac{\nu}{2})} \exp\left\{-\frac{1}{2} \operatorname{tr}\left(\Psi \Sigma^{-1}\right)\right\} \tag{3.3.61}$$

with the degrees of freedom $\nu > n - 1$, the positive definite scale matrix $\Psi \in \mathbb{R}^{n \times n}$ and the multivariate gamma function Γ_n. Recalling that

$$p(y(0) \mid \mu_0, \Sigma_0) = \mathcal{N}(y(0) \mid \mu_0, \Sigma_0),$$

the posterior is found as

$$\begin{aligned} p(\mu_0, \Sigma_0 \mid y(0), x_{[0,T]}, -) &\propto p(y(0) \mid \mu_0, \Sigma_0) p(\mu_0, \Sigma_0) \\ &\propto \mathrm{NIW}\Big(\mu_0, \Sigma_0 \,\Big|\, \tilde{\eta}, \tilde{\lambda}, \tilde{\Psi}, \tilde{\kappa}\Big) \end{aligned} \tag{3.3.62}$$

with

$$\begin{aligned} \tilde{\eta} &= \frac{\lambda\eta + y(0)}{\lambda + 1}, \qquad \tilde{\lambda} = \lambda + 1, \qquad \tilde{\kappa} = \kappa + 1, \\ \tilde{\Psi} &= \Big(\Psi^{-1} + \lambda\tilde{\lambda}^{-1}(y(0) - \eta)(y(0) - \eta)^\top\Big)^{-1}. \end{aligned} \tag{3.3.63}$$

MJP RATES The prior rates are given by a Gamma distribution:

$$\Lambda_{zz'} \overset{\text{i.i.d.}}{\sim} \mathrm{Gam}(s, r) \quad \forall z, z' \in \mathcal{Z},\ z' \neq z, \tag{3.3.64}$$

with the shape $s \in \mathbb{R}_{>0}$ and the rate parameter $r \in \mathbb{R}_{>0}$. The set of all rates is denoted as $\{\Lambda_{zz'}\}$. As detailed in Section 2.1.2.1, the trajectory $z_{[0,T]}$ can be unambiguously expressed in terms of its state sequence $\{z_k\}$ and the corresponding sojourn times $\{s_k\}$, $k = 0, \ldots, K$. The likelihood function

$$p(z_{[0,T]} \mid \{\Lambda_{zz'}\}) = \prod_{z \in \{z_k\}} e^{-\Lambda_z t_z} \prod_{z' \in \mathcal{Z} \setminus z} \Lambda_{zz'}^{n_{zz'}} \tag{3.3.65}$$

with the number of transitions $n_{zz'} = \sum_k \mathbb{1}(z_k = z \wedge z_{k+1} = z')$ and the cumulative sojourn times $t_z = \sum_k \mathbb{1}(z_k = z) s_k$, see Eq. (2.1.17).

As an aside, note that there is a catch for the smoothing case (as opposed to the plain forward simulation of an MJP): in the smoothing setting, one has by definition an upper time limit T. Forward simulation yields tuples of states and sojourn times $\{(z_k, s_k) : k = 0, \ldots, K\}$ until $j_K < T < j_{K+1}$ with the jump times j_k, cf. Section 2.1.2.1. It is only by the properties of the exponential sojourn time distribution that the "remainder" after the last state switch within $[0, T]$ at j_K, i.e. $\Delta s_K := T - j_K$, is also exponentially distributed. For notational convenience, we write in an abuse of notation $s_K \leftarrow \Delta s_K$, that is, we interpret the time between the last jump and the end of the interval as the last sojourn time.

By multiplication with the prior (3.3.64) it is readily checked that this yields a Gamma distribution as the rate posterior:

$$\begin{aligned} p(\Lambda_{zz'} \mid z_{[0,T]}) &\propto p(z_{[0,T]}, - \mid \Lambda_{zz'}) p(\Lambda_{zz'}) \\ &\propto \mathrm{Gam}(\Lambda_{zz'} \mid s + n_{zz'}, r + t_z). \end{aligned} \tag{3.3.66}$$

SDE DRIFT PARAMETERS In the following, we utilize (cf. Eq. (3.1.8)) the shorthand $\Gamma_z = [A(z), b(z)]$. The SSDE parameters Γ_z are specified via a Matrix-Normal (MN) prior

$$\begin{aligned} p(\Gamma_z) &= \mathrm{MN}(\Gamma_z \mid M_z, D_z, K_z) \\ &= \frac{|D_z|^{-\frac{n}{2}} |K_z|^{-\frac{n+1}{2}}}{(2\pi)^{\frac{n(n+1)}{2}}} \exp\left\{-\frac{1}{2} \operatorname{tr}\left((\Gamma_z - M_z)^\top D_z^{-1} (\Gamma_z - M_z) K_z^{-1}\right)\right\}, \end{aligned} \tag{3.3.67}$$

where $D_z = D(z)$ is the SSDE covariance, $M_z \in \mathbb{R}^{n\times(n+1)}$ the location matrix and $K_z \in \mathbb{R}^{(n+1)\times(n+1)}$ the scale matrix. Expressing the conditional Y-posterior via the Radon-Nikodym derivative $G(y_{[0,T]}, z_{[0,T]})$, c.f. Eq. (3.3.42), one can interpret G as the likelihood of the drift parameters, $G(y_{[0,T]}, z_{[0,T]}) = G(y_{[0,T]}, z_{[0,T]} \mid \{\Gamma_z\})$,

$$p(\Gamma_z \,|\, y_{[0,T]}, z_{[0,T]}) \propto G(y_{[0,T]}, z_{[0,T]} \,|\, \{\Gamma_z\}) p(\Gamma_z). \tag{3.3.68}$$

This "likelihood term" can be evaluated approximately by inserting the simulated paths, that is, via the Euler-Maruyama approximation of the SSDE. For the mode z, only those subintervals of $z_{[0,T]}$ contribute in which $z(t) = z$. Accordingly,

$$\begin{aligned} p(\Gamma_z \mid y_{[0,T]}, z_{[0,T]}, -) = \exp\Bigg\{ &\sum_{k\,:\, z(j_k)=z} \int_{j_k}^{j_{k+1}} f(y(s), z)^\top D(z)^{-1} \mathrm{d}y(s) \\ &- \frac{1}{2}\int_{j_k}^{j_{k+1}} f(y(s), z)^\top D(z)^{-1} f(y(s), z)\mathrm{d}s \Bigg\} p(\Gamma_z). \end{aligned}$$

with the jump times $\{j_k\}$, see above and Section 2.1.2.1. Omitting the sum over intervals for readability, we find (for any interval $[j_k, j_{k+1})$) upon inserting the simulated SSDE-path $y_{[0,T]}$

$$\begin{aligned} &\exp\left\{ \int_{j_k}^{j_{k+1}} f(y(s), z)^\top D(z)^{-1} \mathrm{d}y(s) - \frac{1}{2}\int_{t_0}^{t_1} f(y(s), z)^\top D(z)^{-1} f(y(s), z)\mathrm{d}s \right\} \\ &\qquad \approx \exp\left\{ \sum_{l=1}^{L} f^\top(y_l, z) D(z)^{-1} \Delta y_l - \frac{1}{2}\sum_{l=1}^{L} f^\top(y_l, z) D(z)^{-1} f(y_l, z) h \right\}, \end{aligned}$$

where h is time simulation time-step, $s_l = s_{l-1} + h$, the interval boundaries $s_1 = j_k$, $s_L = j_{k+1}$, and $\Delta y_l := y(s_l) - y(s_{l-1})$ the difference of two successive points of the trajectory. Inserting the drift $f(y_l, z) = \Gamma_z \bar{y}_l$, where $\bar{y}_l = \left[y_l^\top, 1_n^\top\right]^\top$ yields

$$\begin{aligned} &\exp\left\{ \sum_{l=1}^{L} f(y_l, z)^\top D(z)^{-1} \Delta y_l - \frac{1}{2}\sum_{l=1}^{L} f(y_l, z)^\top D(z)^{-1} f(y_l, z) h \right\} \\ &= \exp\left\{ -\frac{1}{2}\sum_{l=1}^{L} \left(\frac{\Delta y_l}{\sqrt{h}} - \Gamma_z \bar{y}_l \sqrt{h} \right)^\top D(z)^{-1} \left(\frac{\Delta y_l}{\sqrt{h}} - \Gamma_z \bar{y}_l \sqrt{h} \right) \right. \\ &\qquad \left. + \frac{1}{2}\sum_{l=1}^{L} \Delta y_l^\top D(z)^{-1} \Delta y_l \frac{1}{h} \right\}. \end{aligned} \tag{3.3.69}$$

The last term on the right-hand side may be omitted, as it is independent of Γ_z; the above is hence equivalent to approximating the Radon-Nikodym derivative $G(y_{[0,T]}, z_{[0,T]})$ via a product of Gaussian transition distributions. Making use of the trace function and defining the joint observation vectors

$$\begin{aligned} \Delta y &:= \left[\frac{\Delta y_1}{\sqrt{h}}, \cdots, \frac{\Delta y_L}{\sqrt{h}} \right] \in \mathbb{R}^{n\times L}, \\ \bar{y} &:= \left[\bar{y}_1 \sqrt{h}, \cdots, \bar{y}_1 \sqrt{h} \right] \in \mathbb{R}^{n+1\times L}, \end{aligned}$$

this can be re-formulated as

$$\exp\left\{-\frac{1}{2}\sum_{l=1}^{L}\left(\frac{\Delta y_l}{\sqrt{h}}-\Gamma_z\bar{y}_l\sqrt{h}\right)^{\top}D(z)^{-1}\left(\frac{\Delta y_l}{\sqrt{h}}-\Gamma_z\bar{y}_l\sqrt{h}\right)\right\}$$
$$=\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(\Delta y-\Gamma_z\bar{y}\right)^{\top}D(z)^{-1}\left(\Delta y-\Gamma_z\bar{y}\right)\mathbb{1}_{L\times L}\right)\right\}.$$

Comparison with Eq. (3.3.67) reveals that this expression corresponds to an (un-normalized) MN distribution. Accordingly, we find

$$p(\Gamma_z \mid y_{[0,T]}, z_{[0,T]}, -) \propto \mathrm{MN}(\Delta y \mid \Gamma_z\bar{y}, D_z, \mathbb{1}_{L\times L})p(\Gamma_z) \tag{3.3.70}$$

As it is known that the MN distribution is itself a conjugate prior to the MN likelihood in the above form [96], the sought-after posterior is again MN,

$$p(\Gamma_z \mid y_{[0,T]}, z_{[0,T]}, -) = \mathrm{MN}(\Gamma_z \mid \tilde{M}_z, D_z, \tilde{K}_z) \tag{3.3.71}$$

with posterior hyperparameters

$$\tilde{K}_z = \bar{y}\bar{y}^{\top} + K_z, \qquad \tilde{M}_z = (\Delta y\bar{y}^{\top} + M_zK_z)\tilde{K}_z^{-1}. \tag{3.3.72}$$

Summation over all intervals with $z(j_k) = z$ is straightforward. This derivation also holds for adaptive step-sizes, $s_l = s_{l-1} + h_{l-1}$.

As a caveat, note that this prior does not guarantee stability of the individual modes: it does not impose any constraints on the eigenvalue spectra of the sub-matrices $A(z)$, which determine the asymptotic properties of the diffusion, cf. Section 2.1.2.2. It is known, however, that for switching systems, global stability of the system does *not* require strict intra-mode stability [118]. Additionally, from a practical perspective, the posterior will be likely peaked around stable matrices if conditioned on data from a stable mode - which we expect to observe, as we are interested in locally stable systems in the first place. This is a common assumption also for discrete-time SLDS [119].

SDE DISPERSION The dispersion $Q(z)$ has a special role among the model parameters, as (i) together with the used time step-size, it determines the accuracy of the SDE solver, and (ii) two SDEs with different dispersions are singular with respect to each other [120]. Hence, while a posterior dispersion can be derived along the same lines as the drift parameters (3.3.68), the resulting posteriors may cause instabilities in the solver if Q becomes too large (because of (i)), or it may exhibit poor mixing properties (due to (ii)). Potential solutions to the latter exist, but they are themselves quite involved [84, 120]. As this might be considered a rather exotic issue (sifting through the literature, the reader will find that, quite frequently, the dispersion is assumed to be fixed), we accept this potential drawback for the present purposes.

To still ensure numerical stability, we employ a Metropolis-adapted Langevin algorithm [121]. As the name suggests, this sampling scheme proceeds in two steps:

1. a new state is proposed via Langevin dynamics, utilizing the gradient of the target PDF,
2. the proposed state is accepted (or rejected) via the Metropolis-Hastings algorithm [9].

We simulate an SDE in the space of dispersion matrices via the Euler-Maruyama approximation with step-size $0 < \xi \ll 1$, [121]

$$Q_z^* = Q_z + \xi \partial_{Q_z} \log p(Q_z \mid \{y_{hl}\}) + \sqrt{2\xi}\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \mathbb{1}_{n\times n}), \tag{3.3.73}$$

where $p(Q_z \mid \{y_{hl}\})$ is the approximation of $p(Q_z \mid y_{[0,T]}) \propto G(y_{[0,T]}, z_{[0,T]})p(Q_z)$ on the SDE simulation time grid. As shown above, the approximate density $p(Q_z \mid \{y_{hl}\})$ is equivalent to a product of Gaussian transition distributions $\mathcal{N}(y_l \mid y_{l-1}, D_z h)$, allowing the gradient to be evaluated. The proposed dispersion Q_z^* is then accepted with probability

$$A(Q, Q^*) = \frac{p(Q_z^* \mid y_{[0,T]})q(Q \mid Q^*)}{p(Q_z \mid y_{[0,T]})q(Q^* \mid Q)}, \tag{3.3.74}$$

where q denotes the (Gaussian) proposal density induced by Eq. (3.3.73).

OBSERVATION COVARIANCE Lastly, an IW prior is imposed on the observation covariance Σ_x,

$$p(\Sigma_x) = \mathrm{IW}(\Sigma_x \mid \Psi_x, \lambda_x). \tag{3.3.75}$$

With the Gaussian observations $X_i \sim \mathcal{N}(y_i, \Sigma_x)$, the standard result is

$$p(\Sigma_x \mid x_{[0,T]}) = \mathrm{IW}(\Sigma_x \mid \tilde{\Psi}_x, \tilde{\lambda}_x) \tag{3.3.76}$$

where

$$\tilde{\Psi}_x = \sum_{i=1}^{N} x_i x_i^\top + \Psi_x, \qquad \tilde{\lambda}_x = N + \lambda. \tag{3.3.77}$$

The full Gibbs sampling algorithm is provided in Fig. 3.3.

```
input : observation data $\{t_i, x_i\}_{i=1,\dots,N}$
Initialize $z^0_{[0,T]}, y^0_{[0,T]}, \theta^0$
for $i = 0, \dots, \mathrm{NumSamples}$ do
    Given $z^i_{[0,T]}$, compute $\partial_y \ln \beta$ using Eq. (3.3.36)
    Given $z^i_{[0,T]}$, sample $y^{i+1}_{[0,T]}$ according to Eq. (3.3.27)
    Given $y^{i+1}_{[0,T]}$, compute $p_f$ using Eq. (3.3.51) (equivalently $\tilde{p}_f$ using Eq. (3.3.50))
    Given $y^{i+1}_{[0,T]}$, sample $z^{i+1}_{[0,T]}$ according to Eq. (3.3.55)
    Given $z^{i+1}_{[0,T]}, y^{i+1}_{[0,T]}$, sample model parameters $\theta^{i+1}$
end
```

FIGURE 3.3: Gibbs sampler for MJP-switching stochastic differential equations

3.4 RESULTS

The presented Gibbs sampler is first verified on data generated under the modeling assumption. Subsequently, we apply it to fluorescence data from an inducible gene expression system to infer its latent expression states. Details on the simulations and initializations are provided in Appendix B.4.

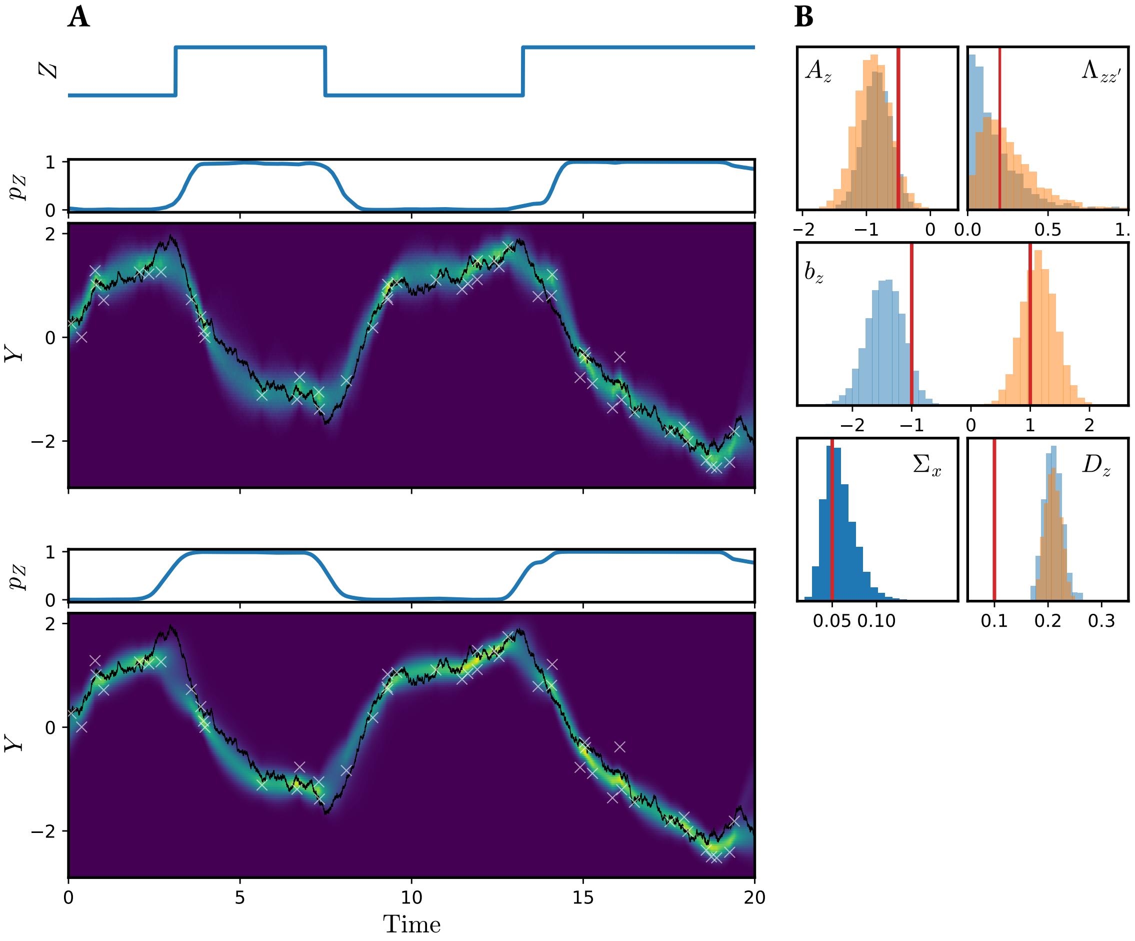


FIGURE 3.4: Model validation on 1D ground-truth data. **A**: system trajectories. Top: ground-truth switching trajectory $z_{[0,T]}$. Middle: empirical posteriors $p(z, t \mid x_{[0,T]})$ and $p(y, t \mid x_{[0,T]})$. Brighter colors indicate higher probability density. Black solid line: ground-truth latent trajectory $y_{[0,T]}$. White crosses: observations. $N_{\text{samples}} = 10000$. Bottom: respective marginals of the solution of the exact PDE (3.2.20) with the ground-truth parameters. **B**: parameter estimates of the drift parameters $A(z)$, $b(z)$, cf. Eq. (3.1.7), the MJP rates $\Lambda(z, z')$, the SDE covariance $D(z)$ and the observation covariance Σ_x. Red lines: ground-truth values. Blue and orange shading indicates the two modes $z = 1, 2$ where applicable.

3.4.1 *Verification on ground-truth data*

1D SYSTEM First, the method is tested on ground-truth data from a one-dimensional two-mode switching system. The mode dynamics are - as specified above - given by time-independent linear drift functions

$$f(y, z, t) = A(z)y + b(z). \tag{3.4.78}$$

We choose $A(z) < 0$ for both z, which makes the individual mode dynamics instances of the well-known Ornstein-Uhlenbeck process [20].

It can be seen in Fig. 3.4 A that the MCMC backward-forward/forward-backward scheme is able to faithfully recover the ground-truth latent trajectories: both $z_{[0,T]}$ and $y_{[0,T]}$ are reproduced with high fidelity. Comparison to the solution of the exact PDE (3.2.20) with ground-truth parameters shows also very good agreement of the time point marginals. Corresponding to the high agreement in the latent sequences, we furthermore obtain accurate Bayesian parameter estimates, see Fig. 3.4 B. All posteriors except the diffusion covariance D_z cover the ground-truth very well. This phenomenon is ascribed to the slow mixing of the Gibbs sampler with respect to $Q(z)$ that was discussed above.

2D SYSTEM To demonstrate the ability of the framework to faithfully reconstruct more complex continuous dynamics, we apply it to a 2D problem in which the continuous component is driven by two counter-rotating vector fields $A(z)$, see Fig. 3.5. Here, we fix the observation covariance Σ_x and the dispersion to the ground-truth value.

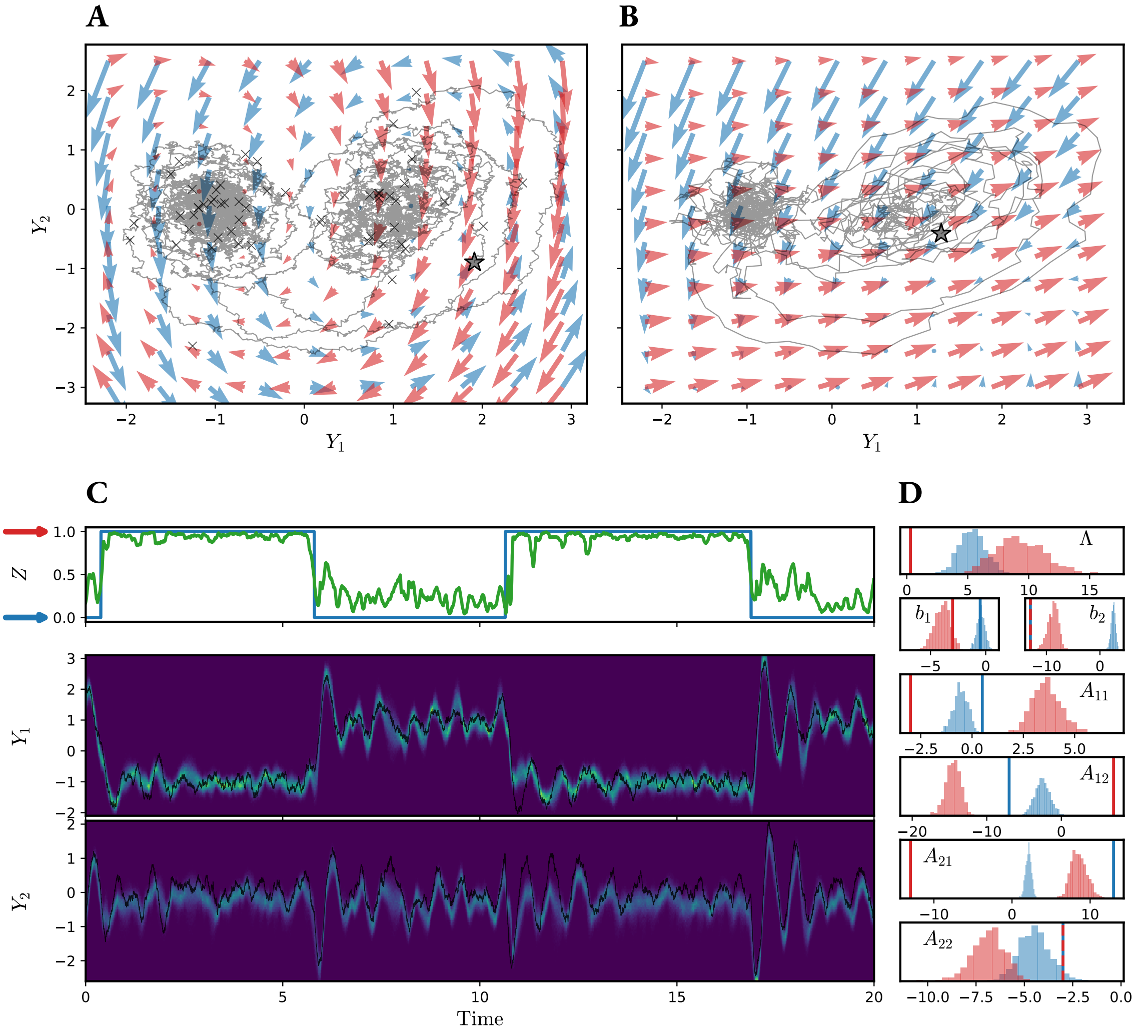


FIGURE 3.5: Model validation on 2D ground-truth data. **A**: flow Eq. (3.4.78) of both modes (red and blue arrows) revolving around $(-1,0)^\top$ and $(1,0)^\top$. Black solid line: ground-truth trajectory $y_{[0,T]}$. Crosses: observed data points; only 20% of all observations are shown to avoid clutter. Star: initial value. **B**: phase space representation of one posterior sample path $y_{[0,T]}$. Arrows represent flows computed with $A(z), b(z)$ averaged over $N = 1000$ samples. **C**: Top: ground-truth trajectory $z_{[0,T]}$ (blue) and empirical posterior $p(z, t \mid x_{[0,T]})$ (green). Middle and Bottom: components of $p(y, t \mid x_{[0,T]})$. Black solid lines as above. **D**: parameter histograms. From top to bottom: exit rates $-\Lambda(z)$ for both modes; individual components $b_1(z), b_2(z)$; individual components $A(z)_{11}, A(z)_{12}, A(z)_{21}, A(z)_{22}$. Lines: ground-truth values. Coloring represents the two separate modes z.

Both the discrete and the continuous dynamics are faithfully recovered as in the 1D example above. We notice, however, that the oscillatory behavior of $y_{[0,T]}$ is reflected in the Z-posterior (Fig. 3.5 C, top). This is plausibilized when plotting the empirical posteriors of $A(z)$ and $b(z)$, Fig. 3.5 B, where we see that the mean vector field of one mode does not complete one "revolution" around the set-point $(1, 0)^\top$. The oscillatory behavior is also reflected in the rate distributions, which strongly overestimate the ground truth. Examined in more detail, the individual histograms over samples of A, b (shown in Fig. 3.5 D) do not cover the ground-truth values well. In particular, we notice that the sign structure of the matrices is not the same as that of the ground-truth. This shows, first, that the MN prior is potentially too unstructured for more complex problems, as it is not straightforward to encode constraints on the eigenspectrum of A. Secondly, comparing the positions of the posterior histograms to the ground-truth values shows that the altered sign structure encompasses A *and* b; this makes sense as these parameters are jointly described by the MN prior. Hence, the posterior distributions also intermingle information about A and b. This shows that while the MN prior on the parameters A, b is sufficient to "get by", it is advisable for more complex systems to employ more richly-structured prior distributions, if possible informed by expert knowledge. The Bayesian parameter updates used here provide a proof of principle, but leave room for improvement. Note also that the runtime may become prohibitive for large systems: for this 2D example, the algorithm ran about two weeks on an Intel Xeon machine.

3.4.2 *Inference of gene-switching dynamics*

To demonstrate the applicability of the Gibbs sampler to real-world problems, we now use the framework to infer the switching dynamics of an inducible gene system. This system was measured in-house in the wet-lab of the Koeppl group: an inducible green fluorescent protein (GFP) was expressed in the eukaryotic model organism *Saccharomyces cerevisiae*. Utilizing a microfluidic platform, gene expression can be induced by a chemical control signal (here: β-estradiol) at arbitrary time points [122]. Expression of the GFP-encoding gene is initiated upon induction. As laid out in the introduction, gene expression generally proceeds in a two-step fashion: first, in the transcription (TX) stage, the gene of interest is read off the DNA and copied into a single-stranded messenger RNA. Second, in the translation (TL) stage, this RNA (coding for a series of amino acids) is utilized to produce the respective protein encoded by the gene. The level of GFP fluorescence is then measured over time through fluorescence microscopy. The TX and TL dynamics are commonly modeled by switching SDEs, the rate parameters of which depend on the stochastic promoter state ("on" vs. "off") of the gene [123]. Here, we infer this latent stochastic promoter state and the GFP level as well as the rate parameters from the data set consisting of noisy microscopy measurements.

While a ground-truth is not available, we can rationalize the inferred promoter activity (see Fig. 3.6 A, top) in the context of the available inducer control signal. Molecular diffusion incurs a certain delay until the promoter gets activated after addition of β-estradiol to the medium. Similarly, upon removal of the inducer, promoter *deactivation* is governed by diffusive export from the cell. Furthermore, the discussed elementary processes of RNA transcription and protein translation also contribute to a delayed activation and deactivation on the protein level with respect to the promoter state. Given these constraints, the inferred expression state as well as GFP level shown in Fig. 3.6 is a plausible reconstruction from a biological point of view.

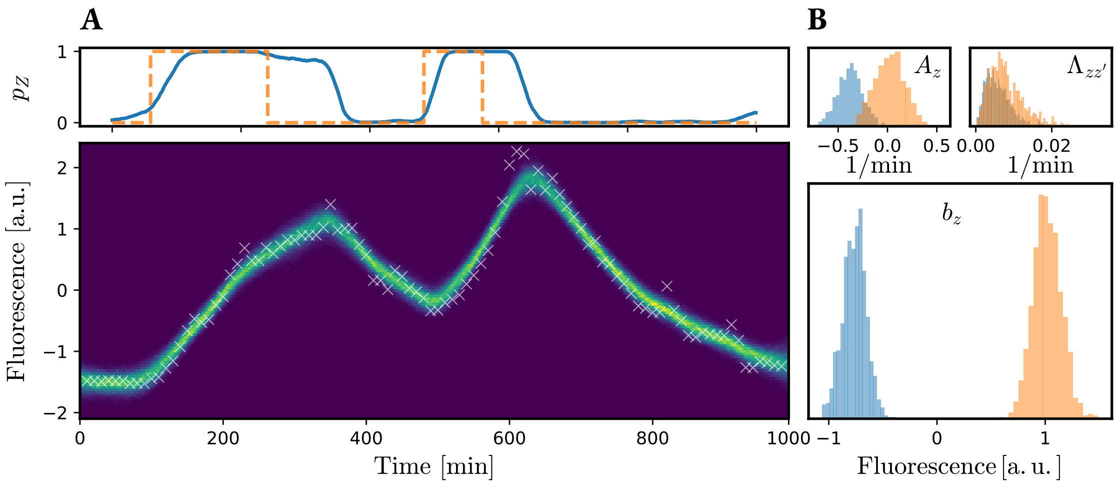


FIGURE 3.6: Inference of promoter states for an inducible gene expression system. **A**: Top: empirical posterior $p(z, t \mid x_{[0,T]})$ (blue) and chemical control (orange; "off", state 1, and "on", state 2). Note that this does not directly correspond to the promoter-"on" and promoter-"off" state, as a the inducer has to diffuse in and out of the cell and its nucleus. Bottom: empirical posterior $p(y, t \mid x_{[0,T]})$. White crosses: observed data. $N_{\text{samples}} = 1000$. **B**: corresponding empirical parameter estimates for the drift parameters $A(z)$ and $b(z)$ as well as the switching rates $\Lambda(z, z')$.

3.5 SUMMARY

We have presented, to the best of our knowledge, the first tractable sampling-based path-space inference scheme for discretely observed continuous-time MJP-SSDEs. The inference algorithm is based on a blocked Gibbs sampler, where we generate sample paths from the exact full conditionals of the switching and diffusion components. To derive these full conditionals, tools from continuous-discrete and fully continuous filtering and smoothing theory are utilized. In contrast to pure latent discrete models, the laid-out framework is also able to reconstruct complex latent continuous dynamics. We furthermore include parameter learning: by the use of conjugate prior distributions, we are able to efficiently sample from the respective posteriors.

The obtained results demonstrate the practical utility of the method: both in the 1D and 2D ground-truth settings, the framework yields accurate results. In the 2D case in particular, it could however also be seen that the parameter learning procedure can be hampered by the structure of the chosen prior distributions. The application to a genetic switching system highlights the potential of the method in the field; it enables the user to learn and utilize an effective model for a transcription-translation system without having to define reaction equations on the molecular level. For an outlook on potential future research, please see Chapter 6.

4

VARIATIONAL INFERENCE FOR HYBRID SYSTEMS

As seen in the last chapter, MCMC sampling from the exact posterior conditional measures of hybrid systems yields accurate results both for the latent trajectories as well as for the parameters. The approach is still computationally expensive, however, resulting in long algorithm runtimes. To address this issue, a variational inference framework is worked out in the following, which only requires passing through the data until convergence instead of one pass *per sample*.

For inference and parameter learning in pure diffusion processes, continuous-time VI frameworks have been developed utilizing, e.g., GPs [87, 88], and general exponential family distributions [23]. Similar methods have also been devised for inference in MJPs [42, 89]. The presented VI framework for MJP-SSDE hybrid systems draws on this previous work and recovers existing diffusion and MJP approximations as special cases.

The approximation is proposed with a specific focus on metastable systems, which remain in distinct, qualitatively different regimes over extended periods of time [124]: a hallmark of such systems is a separation of time scales between the intra- and inter-regime dynamics. Metastable systems are of particular interest for instance in computational structural biology when modeling the conformational switching of complex biomolecules. An extended discussion of metastability in this context will be provided in Chapter 5.

In the following, first the variational problem is stated. Subsequently, the KL divergence between two proper MJP-SSDE processes is derived, based on which the variational approximation is introduced and discussed. After detailing how the optimization problem is solved in practice, the method is evaluated both on ground-truth data and computational biology benchmarks. An implementation of the proposed method is publicly available at `https://git.rwth-aachen.de/bcs/projects/lk/vi-ct-shs.git`.

4.1 THE VARIATIONAL PROBLEM

The MCMC sampling scheme presented in Chapter 3 is based on the characterization of the true posterior - which is determined by the HME (3.2.20) - via a set of ODEs and stochastic PDEs. These equations have to be computed for every sample, rendering this approach computationally expensive. To address this challenge, this chapter takes a VI approach to the same problem: as detailed in Section 2.2.4.2, we aim to find an approximate path measure

$$\mathsf{Q}^* := \mathsf{Q}^*\left((Y_{[0,T]}, Z_{[0,T]}) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]})\right)$$

that is computationally tractable and minimizes the path-wise KL divergence to the exact posterior measure $\mathsf{P}_x := \mathsf{P}\left((Y_{[0,T]}, Z_{[0,T]}) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]}) \mid x_{[0,T]}\right)$:

$$\mathsf{Q}^* := \arg\min_{\mathsf{Q}} \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x] = \arg\min_{\mathsf{Q}} \mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}_x}\right], \tag{4.1.1}$$

where the expectation is taken over Q. The subscript notation is used to discern between the posterior measure (conditioned on the observed data $x_{[0,T]}$) and the prior measure, which will be abbreviated in the following as $\mathsf{P} = \mathsf{P}\left((Y_{[0,T]}, Z_{[0,T]}) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]})\right)$.

As shown in Section 2.2.4.2, the KL divergence in Eq. (4.1.1) can generally be decomposed as

$$\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x] = \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}] - \mathsf{E}[\ln p(x_{[0,T]} \mid Y_{[0,T]})] + \ln p(x_{[0,T]}), \tag{4.1.2}$$

with the expected log-likelihood $\mathsf{E}[\ln p(x_{[0,T]} \mid Y_{[0,T]})] = \sum_{i=1}^{N} \mathsf{E}[\ln p(x(t_i) \mid Y(t_i))]$, allowing to recast the minimization problem Eq. (4.1.1) as a maximization problem over the ELBO

$$\mathsf{L} = \mathsf{E}[\ln p(x_{[0,T]} \mid Y_{[0,T]})] - \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}]. \tag{4.1.3}$$

This expression does not include the marginal log-likelihood $\ln p(x_{[0,T]})$, which is computationally intractable, and can hence be evaluated if the path-wise KL divergence between the variational and the prior measure $\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}]$ is known.

4.2 THE KULLBACK-LEIBLER DIVERGENCE BETWEEN HYBRID PROCESSES

For two hybrid processes of the same class obeying the HME (3.1.3) and inducing measures Q and P, this KL divergence can be derived using Girsanov's theorem, cf. Section 2.1.2.2. Note that the dispersion Q is assumed to be equal in both processes; this is a common assumption (see, e.g., [88]) which is made due to the mutual singularity of two measures with different dispersions, cf. the discussion of the dispersion sampling algorithm in Chapter 3. It is generally possible to relax this assumption [24], but for the present purposes, we stick with the simpler case and (as before) assume $Q = Q(z)$. To obtain the KL divergence, first note that we can decompose

$$\mathsf{P}\left((Y_{[0,T]}, Z_{[0,T]}) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]})\right) = \mathsf{P}_{Y|Z}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})\mathsf{P}_Z(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}),$$

and analogously for Q, where we introduce subscripts for concision in the following. We can therefore write the expectation as

$$\mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}}\right] = \mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}_Z}{\mathrm{d}\mathsf{P}_Z} + \ln \frac{\mathrm{d}\mathsf{Q}_{Y|Z}}{\mathrm{d}\mathsf{P}_{Y|Z}}\right]. \tag{4.2.4}$$

The first term on the right-hand side is the contribution of the MJP components. The second term is the SSDE contribution: recalling the argument given in Section 3.3.1 that the conditional measure $\mathsf{P}_{Y|Z}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})$ can be understood as a succession of conventional, differently parameterized SDEs, it is readily understood that this term is an expectation over a sum of SDE-contributions.

MJP CONTRIBUTION Utilizing the Radon-Nikodym derivative provided in Chapter 3, cf. Eq. (2.1.19), we find

$$\begin{aligned}\mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}_Z}{\mathrm{d}\mathsf{P}_Z}\right] = \mathsf{E}\Bigg[\int_0^T &\Lambda(Z(s), s) - \tilde{\Lambda}(Z(s), s)\mathrm{d}s \\ &+ \sum_{s \in j_{[0,T]}} \ln\left(\frac{\tilde{\Lambda}(Z(s))q(Z(s), s \mid Z(s^-))}{\Lambda(Z(s))p(Z(s), s \mid Z(s^-))}\right)\Bigg] + \mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}_{Z(0)}}{\mathrm{d}\mathsf{P}_{Z(0)}}\right],\end{aligned}$$

where the last term represents the contribution of the initial distributions,

$$\mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}_{Z(0)}}{\mathrm{d}\mathsf{P}_{Z(0)}}\right] =: \mathsf{D}_{\mathrm{KL}}[\mathsf{Q}_Z^0 \mid\mid \mathsf{P}_Z^0].$$

We aim to replace the unwieldy sum over jumps with an expectation. To that end, recall that

$$q(Z(s) = z', s \mid Z(s^-) = z) = \frac{\tilde{\Lambda}(z, z', s)}{\tilde{\Lambda}(z, s)},$$

and similarly for p, see Section 2.1.2.1.

Now we discretize, $t_k \in \{0, h, 2h, \ldots, K \cdot h = T\}$, allowing us to write

$$\begin{aligned}\mathsf{E}\left[\sum_{s \in j_{[0,T]}} \ln\left(\frac{\tilde{\Lambda}(Z(s), Z(s^-), s)}{\Lambda(Z(s), Z(s^-), s)}\right)\right] = \\ \sum_k \mathsf{E}\left[\mathsf{E}\left[\sum_{s \in j_{[t_k, t_k+h]}} \ln\left(\frac{\tilde{\Lambda}(Z(s), Z(s^-), s)}{\Lambda(Z(s), Z(s^-), s)}\right)\middle| Z(t_k)\right]\right].\end{aligned}$$

By definition, we have

$$q(Z(t+h)=z', Z(t)=z, t) = \delta_{zz'} + \tilde{\Lambda}(z, z', t)h + o(h)$$

and analogously for p; as we will let $h \to 0$, we can confine the (generally infinite) sum to only one transition and neglect higher order terms:

$$\begin{aligned}\mathsf{E}\left[\sum_{s\in j_{[t_k, t_k+h]}} \ln\left(\frac{\tilde{\Lambda}(Z(s), Z(s^-), s)}{\Lambda(Z(s), Z(s^-), s)}\right)\middle| Z(t_k)\right] \\ = \sum_{z'\in\mathcal{Z}\setminus Z(t_k)} \tilde{\Lambda}(Z(t_k), z', t_k)\ln\left(\frac{\tilde{\Lambda}(Z(t_k), z', t_k)}{\Lambda(Z(t_k), z', t_k)}\right) h + o(h).\end{aligned}$$

Consequentially, we can replace the outer sum with an integral and obtain

$$\begin{aligned}\mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_{Z(0)}}{\mathrm{d}\mathsf{P}_{Z(0)}}\right] &= \mathsf{E}\left[\int_0^T \Lambda(Z(t), t) - \tilde{\Lambda}(Z(t), t)\mathrm{d}t\right] \\ &+ \int_0^T \sum_{z\in\mathcal{Z}} q(z,t) \sum_{z'\in\mathcal{Z}\setminus z} \tilde{\Lambda}(z, z', t)\ln\left(\frac{\tilde{\Lambda}(z, z', t)}{\Lambda(z, z', t)}\right)\mathrm{d}t + \mathsf{D}_{\mathrm{KL}}\left[\mathsf{Q}_Z^0 \mid\mid \mathsf{P}_Z^0\right] \\ &= \int_0^T \sum_{z\in\mathcal{Z}} q(z,t)\Bigg(\Lambda(z,t) - \tilde{\Lambda}(z,t) \\ &+ \sum_{z'\in\mathcal{Z}\setminus z} \tilde{\Lambda}(z, z', t)\ln\left(\frac{\tilde{\Lambda}(z, z', t)}{\Lambda(z, z', t)}\right)\Bigg)\mathrm{d}t + \mathsf{D}_{\mathrm{KL}}\left[\mathsf{Q}_Z^0 \mid\mid \mathsf{P}_Z^0\right].\end{aligned} \tag{4.2.5}$$

SSDE CONTRIBUTION This contribution is straightforwardly derived by inserting the Radon-Nikodym derivative between two SDEs, cf. Eq. (2.1.26):

$$\begin{aligned}\mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_{Y|Z}}{\mathrm{d}\mathsf{P}_{Y|Z}}\right] &= \mathsf{E}\Bigg[\int_0^T (f(Y(t), Z(t), t) - g(Y(t), Z(t), t))^\top D^{-1}(Z(t))\mathrm{d}W_{\mathsf{P}}(t) \\ &-\frac{1}{2}\int_0^T \|g(Y(t), Z(t), t) - f(Y(t), Z(t), t)\|^2_{D^{-1}}\mathrm{d}t\Bigg] + \mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_{Y(0)|Z}}{\mathrm{d}\mathsf{P}_{Y(0)|Z}}\right],\end{aligned}$$

with, as above, the last term representing the KL divergence of the initial distributions,

$$\mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_{Y(0)|Z}}{\mathrm{d}\mathsf{P}_{Y(0)|Z}}\right] =: \mathsf{D}_{\mathrm{KL}}\left[\mathsf{Q}^0_{Y|Z} \mid\mid \mathsf{P}^0_{Y|Z}\right],$$

and the shorthand $\|f\|^2_{D^{-1}} := f^\top D^{-1} f$. Due to Girsanov's theorem Eq. (2.1.27), we have

$$\mathrm{d}W_{\mathsf{P}}(t) = \mathrm{d}W_{\mathsf{Q}}(t) + (f(Y(t), Z(t), t) - g(Y(t), Z(t), t))\mathrm{d}t,$$

which allows us to rewrite

$$\begin{aligned}\int_0^T & \left(f(Y(t), Z(t), t) - g(Y(t), Z(t), t)\right)^\top D^{-1}(Z(t))\mathrm{d}W_{\mathsf{P}}(t) \\ &= \int_0^T \left(f(Y(t), Z(t), t) - g(Y(t), Z(t), t)\right)^\top D^{-1}(Z(t))\mathrm{d}W_{\mathsf{Q}}(t) \\ &\quad + \int_0^T \|f(Y(t), z(t), t) - g(Y(t), z(t), t)\|^2_{D^{-1}}\mathrm{d}t.\end{aligned}$$

The Q-Brownian motion vanishes under the expectation (cf. Section 2.1.2.2), and we are left with

$$\mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{Q}_{Y|Z}}{\mathrm{d}\mathsf{P}_{Y|Z}}\right] = \int_0^T \frac{1}{2}\sum_{z\in\mathcal{Z}} q(z,t)\,\mathsf{E}\left[\|g(Y(t), z, t) - f(Y(t), z, t)\|^2_{D^{-1}}\big|z\right]\mathrm{d}t \tag{4.2.6}$$

The full KL divergence Eq. (4.2.4) with Eqs. (4.2.5) and (4.2.6) finally reads

$$\begin{aligned}\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}] = \mathsf{D}_{\mathrm{KL}}[\mathsf{Q}^0 \mid\mid \mathsf{P}^0] + \int_0^T \frac{1}{2}\,\mathsf{E}\left[\|f(Y(t), Z(t), t) - g(Y(t), Z(t), t)\|^2_{D^{-1}}\right] \\ + \sum_{z\in\mathcal{Z}} q(z,t)\left\{(\Lambda(z,t) - \tilde{\Lambda}(z,t)) + \sum_{z'\in\mathcal{Z}\setminus z}\tilde{\Lambda}(z,z',t)\ln\left(\frac{\tilde{\Lambda}(z,z',t)}{\Lambda(z,z',t)}\right)\right\}\mathrm{d}t,\end{aligned} \tag{4.2.7}$$

where the two individual initial contributions are summarized as $\mathsf{D}_{\mathrm{KL}}[\mathsf{Q}^0 \mid\mid \mathsf{P}^0]$. As an aside, note that the same result can be obtained by discretizing the measures Q and P upfront and using a limiting procedure, as done, e.g., in [42]. This alternate derivation is provided in Appendix C.1, which also demonstrates clearly why the KL between two processes with different dispersions Q would diverge without further assumptions. Notably, in the absence of any coupling between $Z(t)$ and $Y(t)$, i.e., $f(y, z, t) = f(y, t)$, Eq. (4.2.7) reduces to the sum of the known individual path-wise KL divergences for diffusion processes and MJPs [42, 87].

With Eq. (4.2.7), we can write the full ELBO as an integral expression:

$$\begin{aligned}\mathsf{L} = &- \mathsf{D}_{\mathrm{KL}}[\mathsf{Q}^0 \mid\mid \mathsf{P}^0] \\ &+ \int_0^T \mathsf{E}\Big[\sum_{i=1}^N \ln p(x_i \mid Y(t_i)) - \frac{1}{2}\|g(Y(t), Z(t), t) - f(Y(t), Z(t), t)\|^2_{D^{-1}}\Big] \\ &- \sum_{z\in\mathcal{Z}} q(z,t)\left\{(\Lambda(z,t) - \tilde{\Lambda}(z,t)) + \sum_{z'\in\mathcal{Z}\setminus z}\tilde{\Lambda}(z,z',t)\ln\left(\frac{\tilde{\Lambda}(z,z',t)}{\Lambda(z,z',t)}\right)\right\}\mathrm{d}t.\end{aligned} \tag{4.2.8}$$

As in the following, we will make frequent use of the *integrand* of L, we define

$$\mathsf{L} =: \int_0^T \ell_{\mathsf{L}}\mathrm{d}t. \tag{4.2.9}$$

4.3 STRUCTURED MEAN-FIELD VARIATIONAL INFERENCE

The unconstrained optimization problem Eq. (4.1.1) is solved by the true, but computationally intractable posterior distribution P_x, cf. Section 2.2.4.2. To arrive at *tractable* expressions, we further need to restrict the class of admissible variational processes.

The conventional mean-field approach to this would read

$$\mathsf{Q}\left((Y_{[0,T]}, Z_{[0,T]}) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]})\right) = \mathsf{Q}(Y_{[0,T]} \in \mathrm{d}y_{[0,T]})\mathsf{Q}(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}).$$

In the following, we will depart from this standard assumption and instead retain a dependency between the Y- and Z-processes,

$$\mathsf{Q}\left((Y_{[0,T]}, Z_{[0,T]}) \in \mathrm{d}(y_{[0,T]}, z_{[0,T]})\right) = \mathsf{Q}_Y(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})\mathsf{Q}_Z(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}),$$

where we again have introduced subscripts to avoid ambiguity between the different measures - the same will also be done for P, as above. To maintain tractability, we will impose additional structure on the conditional measure $\mathsf{Q}_Y(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]})$. We do so by choosing a variational process that consists of a *set* of SDEs:

$$\begin{aligned} Z &\sim \mathrm{MJP}\left(\tilde{\Lambda}(z, z', t)\right) =: \mathsf{Q}_Z, \\ Y^z(t) &\sim \mathrm{SDE}\left(g(y, z, t), Q(y, z, t)\right) =: \mathsf{Q}_Y^z \text{ for } z \in \mathcal{Z}\text{, where} \\ \mathrm{d}Y^z(t) &= g(Y(t), z, t)\mathrm{d}t + Q(Y(t), z, t)\mathrm{d}W(t), \end{aligned} \tag{4.3.10}$$

where each SDE Y^z induces a measure Q_Y^z. Therefore, this constitutes a mixture of SDEs rather than a proper SSDE and corresponds to a *structured* mean-field factorization of the conditional measure,

$$\begin{aligned} &\mathsf{Q}_Y(Y_{[0,T]} \in \mathrm{d}y_{[0,T]} \mid z_{[0,T]}) \\ &\qquad = \prod_{z \in \mathcal{Z}} \prod_{k \in j_{[0,T]}} \left(\mathsf{Q}_Y^z\left(Y^z_{[j_k, j_{k+1})} \in \mathrm{d}y_{[j_k, j_{k+1})}\right)\right)^{\mathbb{1}(z_{[j_k, j_{k+1})} = z)} \end{aligned} \tag{4.3.11}$$

with the product over all jumps occurring in $z_{[0,T]}$ (with respective jump times j_k, cf. Section 2.1.2.1) and all possible components z: in analogy to discrete-time mixture models [9], the piece-wise constant $z_{[j_k, j_{k+1})} = z$ selects the mixture component Y^z. With this, the full structured mean-field ansatz reads

$$\begin{aligned} &\mathsf{Q}(Y_{[0,T]}, Z_{[0,T]} \in \mathrm{d}(y_{[0,T]}, z_{[0,T]})) \\ &\qquad = \mathsf{Q}_Z(Z_{[0,T]} \in \mathrm{d}z_{[0,T]}) \prod_{z \in \mathcal{Z}} \prod_{k \in j_{[0,T]}} \left(\mathsf{Q}_Y^z\left(Y^z_{[0,T]} \in \mathrm{d}y_{[0,T]}\right)\right)^{\mathbb{1}(z_{[j_k, j_{k+1})} = z)}. \end{aligned} \tag{4.3.12}$$

We choose g to be linear for all z; thus, Eq. (4.3.10) corresponds to a set of GPs,

$$Y^z_{[0,T]} \sim \mathrm{GP}\left(m_z, C_z\right) \; \forall z \in \mathcal{Z}.$$

The proposed approximation can accordingly be understood as a *mixture* of GPs. Note that for the present case, the mean functions m_z and covariance function C_z could be computed explicitly, but this is of no interest here; see [20, 125] for an in-depth review of GPs and their relation to SDEs.

The KL divergence between the induced measure Q and the true MJP-SSDE measure P can then be decomposed as follows:

$$\begin{aligned}\mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}}\right] &= \mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_Z}{\mathrm{d}\mathsf{P}_Z}\right] \\ &+ \mathsf{E}\left[\mathsf{E}\left[\sum_{k\in j_{[0,T]}}\sum_{z\in\mathcal{Z}}\mathbb{1}(Z_{[j_k,j_{k+1})}=z)\ln\frac{\mathrm{d}\mathsf{Q}^z_{Y^z_{[j_k,j_{k+1})}}}{\mathrm{d}\mathsf{P}_{Y_{[j_k,j_{k+1})}|Z_{[j_k,j_{k+1}]}}}\,\middle|\, Z_{[0,T]}\right]\right]\end{aligned}$$

where

$$\begin{aligned}\mathsf{Q}^z_{Y^z_{[j_k,j_{k+1})}} &= \mathsf{Q}^z_Y\left(Y^z_{[j_k,j_{k+1})}\in \mathrm{d}y_{[j_k,j_{k+1})}\right), \\ \mathsf{P}_{Y_{[j_k,j_{k+1})}|Z_{[j_k,j_{k+1})}} &= \mathsf{P}_Y\left(Y_{[j_k,j_{k+1})}\in\mathrm{d}y_{[j_k,j_{k+1})}\mid z_{[j_k,j_{k+1})}\right).\end{aligned}$$

The first term is just the KL divergence between two MJPs that was already provided above, see Eq. (4.2.5). The second term is more involved. To render it tractable via a time point-wise expectation, note that each summand can be expressed via Eq. (4.2.6)

$$\begin{aligned}&\mathsf{E}\left[\mathbb{1}(z_{[j_k,j_{k+1})}=z)\ln\frac{\mathrm{d}\mathsf{Q}^z_{Y^z_{[j_k,j_{k+1})}}}{\mathrm{d}\mathsf{P}_{Y_{[j_k,j_{k+1})}|Z_{[j_k,j_{k+1}]}}}\,\middle|\, Z_{[0,T]}=z_{[0,T]}\right] \\ &\quad= \frac{1}{2}\int_{j_k}^{j_{k_1}}\mathbb{1}(z(t)=z)\,\mathsf{E}\left[\|f(Y(t),z,t)-g(Y(t),z,t)\|^2_{D^{-1}}\mid z\right]\mathrm{d}t \\ &\quad+ \mathsf{E}\left[\mathbb{1}(z_{[j_k,j_{k+1})}=z)\ln\frac{\mathsf{Q}^{z,j_k}_Y}{\mathsf{P}^{j_k}_Y}\right].\end{aligned}$$

Crucially, each of these summands comprises an initial KL, $\mathsf{D}_{\mathrm{KL}}\left[\mathsf{Q}^{z,j_k}_Y \mid\mid \mathsf{P}^{j_k}_Y\right]$, which is not computable without knowing the true measure P - the evaluation of which we are trying to avoid with the variational approach. Concatenating all inter-jump intervals in the inner expectation, we can write

$$\begin{aligned}\mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}}\right] &= \mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_Z}{\mathrm{d}\mathsf{P}_Z}\right] \\ &+ \mathsf{E}\left[\mathsf{E}\left[\sum_{z\in\mathcal{Z}}\frac{1}{2}\int_0^T\mathbb{1}(Z(t)=z)\,\mathsf{E}\left[\|f(Y(t),z,t)-g(Y(t),z,t)\|^2_{D^{-1}}\mid z\right]\mathrm{d}t\,\middle|\, Z_{[0,T]}\right]\right] \\ &+ \mathsf{E}\left[\sum_{z\in j_{[0,T]}}\mathbb{1}(z_{[j_k,j_{k+1})}=z)\,\mathsf{E}\left[\ln\frac{\mathsf{Q}^{z,j_k}_Y}{\mathsf{P}^{j_k}_Y}\,\middle|\, Z_{[0,T]}\right]\right].\end{aligned}$$

We now approximate this by omitting the intractable initial KL contributions:

$$\begin{aligned}\mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}}{\mathrm{d}\mathsf{P}}\right] &\approx \mathsf{E}\left[\ln\frac{\mathrm{d}\mathsf{Q}_Z}{\mathrm{d}\mathsf{P}_Z}\right] \\ &+ \frac{1}{2}\int_0^T\mathsf{E}\left[\|f(Y(t),Z(t),t)-g(Y(t),Z(t),t)\|^2_{D^{-1}}\right]\mathrm{d}t. \qquad (4.3.13)\end{aligned}$$

This approximation is applicable in a certain regime, namely, under a separation of time scales: the relaxation of the diffusion Y needs to be much faster than the switching of the Z-process. The relative contributions of the "initial" distributions at jump times decrease as this separation of time scales increases; the fewer the expected jumps in a fixed time interval, the smaller the omitted contribution, and the initial mismatch in the diffusive part of the KL between the true and approximate processes decays as the true process reaches the local steady state. Hence, Eq. (4.3.13) is applicable to metastable systems, which, by definition, transition between different regimes and exhibit a separation of time scales between the intra-mode diffusive dynamics and the inter-mode transitions. Notice that the approximation abolishes the lower-bound property of the so-defined ELBO; a similar approach has been taken e.g. in [97]. While this is certainly undesirable on a conceptual level, it still provides a valuable tool from an engineering perspective, which needs however to be applied with the conditions described above in mind.

The structure of Q, Eq. (4.3.10), entails separate constraints on the dynamics of the processes corresponding to Q_Z and Q_Y^z. These constraints will be formulated on the density level: denote, as in the previous chapter, with $p(y, z, t \mid x_{[0,T]})$ the joint posterior density at time point t, and with $q_Z(z, t), q_Y^z(y, t)$ the corresponding variational densities. Note that the structured mean-field ansatz Eq. (4.3.13) can of course also be expressed on the density level: we can write

$$\begin{aligned} p(y, z, t \mid x_{[0,T]}) &= p(z, t \mid x_{[0,T]}) \cdot p(y, t \mid z, t, x_{[0,T]}) \\ &\approx q_Z(z, t) \cdot q_Y^z(y, t) =: q(y, z, t). \end{aligned} \tag{4.3.14}$$

The exact conditional $p(y, t \mid z, t, x_{[0,T]})$ - which does not generally have a simple parametric form - is approximated by one distribution per mode, $q_Y^z(y, t)$. In the time point-wise mixture distribution $q(y, z, t)$, the densities $q_Y^z(y, t)$ constitute the mixture components and $q_Z(z, t)$ the mixture weights. It is also from this representation immediately clear that omitting the Z-dependency $q_Y^z(y, t) = q_Y(y, t)$ recovers the standard mean-field ansatz [98].

As Q_Z is assumed to be an MJP, the marginal $q_Z(z, t)$ obeys a master equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} q_Z(z, t) = \sum_{z' \in \mathcal{Z} \setminus z} \tilde{\Lambda}(z', z, t) q_Z(z', t) - \tilde{\Lambda}(z, t) q_Z(z, t), \ \forall z \in \mathcal{Z}. \tag{4.3.15}$$

Furthermore, the variational factors $q_Y^z(y, t)$ each follow a FPE with linear variational drift $g(y, z, t) = A_q(z, t) y + b_q(z, t)$ for every mode z individually:

$$\partial_t q_Y^z(y, t) = -\sum_{i=1}^{n} \partial_{y_i} \{g_i(y, z, t) q_Y^z(y, t)\} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_{y_i} \partial_{y_j} \{D_{ij} q_Y^z(y, t)\}. \tag{4.3.16}$$

For linear SDEs, this is solved by a time-dependent Gaussian distribution [20]

$$q_Y^z(y, t) = \mathcal{N}(y \mid \mu(z, t), \Sigma(z, t)),$$

where the dynamics of the parameters are described by two ODEs for all $z \in \mathcal{Z}$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \mu(z, t) &= A_q(z, t) \mu(z, t) + b_q(z, t), \\ \frac{\mathrm{d}}{\mathrm{d}t} \Sigma(z, t) &= A_q(z, t) \Sigma(z, t) + \Sigma(z, t) A_q^\top(z, t) + D. \end{aligned} \tag{4.3.17}$$

The constraints (4.3.15) and (4.3.17) are included into the objective Eq. (4.1.3) via Lagrange multiplier functions, cf. Section 2.2.4.2. Defining the multipliers $\lambda(z,t)$, $\Psi(z,t)$, $\nu(z,t)$ for the variational mean $\mu(z,t)$, covariance $\Sigma(z,t)$ and variational rates $\tilde{\Lambda}(z,z',t)$ respectively, the constrained optimization problem is converted into an unconstrained one:

$$\left.\begin{array}{l}\arg\max_{\mu,\Sigma,\tilde{\Lambda}} \mathsf{L}[\mu,\Sigma,\tilde{\Lambda}] \\ \text{s.t. Eqs. (4.3.15) and (4.3.17)}\end{array}\right\} \longrightarrow \quad \underset{\mu,\Sigma,\tilde{\Lambda}}{\arg\max}\, \mathscr{L}[\mu,\Sigma,\tilde{\Lambda},\lambda,\Psi,\nu],$$

where the full Lagrangian $\mathscr{L}$ reads

$$\begin{aligned}\mathscr{L} = \mathsf{L} + \int_0^T \sum_{z\in\mathcal{Z}} \Bigg[& \lambda^\top(z,t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\mu(z,t) - (A(z,t)\mu(z,t) + b(z,t))\right) \\ & + \operatorname{tr}\left\{\Psi^\top(z,t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\Sigma(z,t) - \big(A(z,t)\Sigma(z,t) + \Sigma(z,t)A^\top(z,t) + D\big)\right)\right\} \\ & + \nu(z,t)\left(\frac{\mathrm{d}}{\mathrm{d}t}q_Z(z,t) - \sum_{z'\in\mathcal{Z}} \tilde{\Lambda}_{z'z}(t) q_Z(z',t)\right)\Bigg] \,\mathrm{d}t =: \int_0^T \ell_{\mathsf{L}}(t) + \ell_{\mathrm{c}}(t)\mathrm{d}t,\end{aligned} \tag{4.3.18}$$

where we defined the integrand of the constraint contributions, $\ell_{\mathrm{c}}(t)$.

MODELING ASSUMPTIONS The same modeling assumptions apply as for the Gibbs sampler in Chapter 3; in particular, the prior drift function f is assumed to be linear and time-invariant,

$$f(y,z,t) = A(z)y + b(z).$$

As defined above, the variational diffusion is also assumed to be linear, but not necessarily time-invariant,

$$g(y,z,t) = A_q(z,t)y + b_q(z,t).$$

The observations are again given as

$$\begin{aligned} X_i = X(t_i) &= Y(t_i) + \zeta, \\ \zeta &\sim \mathcal{N}(0,\Sigma_x). \end{aligned} \tag{4.3.19}$$

Notice that with both drift functions being linear, we can explicitly evaluate the expectation within the ELBO: defining $\bar{A}(z,t) := A_q(z,t) - A(z)$ and $\bar{b}$ analogously, we find

$$\begin{aligned} & \mathsf{E}\left[\|f(Y(t),Z(t),t) - g(Y(t),Z(t),t)\|^2_{D^{-1}}\right] \\ &\quad = \mathsf{E}\left[(\bar{A}(Z(t),t)Y(t) + \bar{b}(Z(t),t))^\top D^{-1}(Z(t))(\bar{A}(Z(t),t)Y(t) + \bar{b}(Z(t),t))\right] \\ &\quad = \sum_{z\in\mathcal{Z}} q_Z(z,t)\left\{\operatorname{tr}\{\bar{A}(z,t)^\top D^{-1}(z)\bar{A}(z,t)\Sigma(z,t)\}\right. \\ &\qquad\quad \left. + (\bar{A}(z,t)\mu(z,t) + \bar{b}(z,t))^\top D^{-1}(z)(\bar{A}(z,t)\mu(z,t) + \bar{b}(z,t))\right\}. \end{aligned}$$

4.4 OPTIMIZING THE VARIATIONAL DISTRIBUTIONS

The structured mean-field assumption allows us to maximize Eq. (4.3.18) individually with respect to $q_Z(z,t)$ and $q_Y^z(y,t)$ [10]. With the introduced parameterization, the optimization problem translates to finding the optimal variational factors

$$q_Z^* := \arg\max_{q_Z} \mathscr{L}, \qquad \mu^* := \arg\max_{\mu} \mathscr{L}, \qquad \Sigma^* := \arg\max_{\Sigma} \mathscr{L},$$

and analogously the parameters A_q^*, b_q^*, $\tilde{\Lambda}^*$, ϕ^*, where ϕ summarizes the variational initial conditions. Pontryagin's maximum principle [126] requires each optimal variational factor $\varphi \in \{q_Z, \mu, \Sigma\}$ to fulfil (i) the respective constraint equations (4.3.15) and (4.3.17) as well as (ii) the Euler-Lagrange (EL) equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\partial_{\dot{\varphi}}\ell = \partial_{\varphi}\ell,$$

with $\ell(t) = \ell_{\mathsf{L}}(t) + \ell_{\mathrm{c}}(t)$ the integrand of the Lagrangian $\mathscr{L}$, cf. Eq. (4.3.18). This translates to an ODE for each Lagrange multiplier function.

4.4.1 *Optimizing the variational switching process*

Firstly, the EL equation with respect to $q_Z(z,t)$ reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\nu(z,t) = \partial_{q_Z(z,t)}\ell_{\mathsf{L}}(t) - \sum_{z'\in\mathcal{Z}\setminus z} \tilde{\Lambda}(z,z',t)\nu(z',t) + \tilde{\Lambda}(z,t)\nu(z,t), \tag{4.4.20}$$

where it is straightforwardly derived that

$$\begin{aligned}\partial_{q_Z(z,t)}\ell_{\mathsf{L}}(t) = -&\sum_{z'\in\mathcal{Z}\setminus z}\left[\tilde{\Lambda}_{zz'}(t)\left(\ln\frac{\tilde{\Lambda}_{zz'}(t)}{\Lambda_{zz'}} - 1 + \nu(z',t) - \nu(z,t)\right) + \Lambda_{zz'}\right]\\ &- \mathsf{E}\left[\|f(Y(t),z,t) - g(Y(t),z,t)\|^2_{D^{-1}}\big|z\right]\\ &+ \sum_{i}^{N}\mathsf{E}\left[\ln p(x(t_i)\mid Y(t_i))\mid z\right]\delta(t-t_i).\end{aligned}$$

In an abuse of notation, the delta function was used outside the integral for compactness. Due to these delta-contributions, the evolution equation for the Lagrange multiplier $\nu(z,t)$ Eq. (4.4.20) is an impulsive differential equation [105] which can - similar to exact posterior inference, see Section 3.2 - be solved piece-wise by integrating the ODE backwards between the discontinuities (starting at $\nu(z,T) = 0$) and applying reset conditions at the integration boundaries:

$$\nu(z,t_i) = \mathsf{E}\left[\ln p(x(t_i)\mid Y(t_i))\middle| z\right] + \nu(z,t_i^-), \tag{4.4.21}$$

with $\nu(z,t_i^-) := \lim_{h\searrow 0}\nu(z,t_i - h)$. For Gaussian observation noise, we have

$$\begin{aligned}\mathsf{E}[\ln p(x(t_i)\mid Y(t_i))|z] = -\frac{1}{2}&\{n\ln(2\pi) + \ln|\Sigma_x|\\ &+(x(t_i) - \mu(z,t_i))^\top\Sigma_x^{-1}(x(t_i) - \mu(z,t_i)) + \mathrm{tr}\{\Sigma_x^{-1}\Sigma(z,t_i)\}\}.\end{aligned} \tag{4.4.22}$$

With the expectation over the drift functions already evaluated above, we have

$$\begin{aligned}\partial_{q_Z(z,t)}\ell_{\mathsf{L}}(t) = &-\operatorname{tr}\{\bar{A}(z,t)^\top D^{-1}\bar{A}(z,t)\Sigma(z,t)\}\\ &-(\bar{A}(z,t)\mu(z,t)+\bar{b}(z,t))^\top D^{-1}(\bar{A}(z,t)\mu(z,t)+\bar{b}(z,t))\\ &-\sum_{z'\in\mathcal{Z}\setminus z}\left[\tilde{\Lambda}_{zz'}(t)\left(\ln\frac{\tilde{\Lambda}_{zz'}(t)}{\Lambda_{zz'}}-1+\nu(z',t)-\nu(z,t)\right)+\Lambda_{zz'}\right]\\ &+\sum_{i=1}^{N}\mathsf{E}\left[\ln p(x(t_i)\mid Y(t_i))\middle|\, z\right]\delta(t-t_i).\end{aligned}$$

4.4.2 *Optimizing the variational diffusion processes*

Secondly, the EL equation has to hold separately for both Gaussian parameters $\mu(z,t)$ and $\Sigma(z,t)$. This yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\lambda(z,t) &= \partial_{\mu(z,t)}\ell_{\mathsf{L}}(t) - A_q^\top(z,t)\lambda(z,t),\\ \frac{\mathrm{d}}{\mathrm{d}t}\Psi(z,t) &= \partial_{\Sigma(z,t)}\ell_{\mathsf{L}}(t) - A_q^\top(z,t)\Psi(z,t) - \Psi(z,t)A_q(z,t).\end{aligned}\tag{4.4.23}$$

We find the gradients as

$$\begin{aligned}\partial_{\mu(z,t)}\ell_{\mathsf{L}}(t) &= -q_Z(z,t)\left(\bar{A}(z,t)^\top D^{-1}\bar{A}(z,t)\mu(z,t)+\bar{A}(z,t)^\top D^{-1}\bar{b}(z,t)\right)\\ &\quad+\sum_{i=1}^{N}q_Z(z,t_i)\Sigma_x^{-1}(x(t_i)-\mu(z,t_i))\delta(t-t_i),\\ \partial_{\Sigma(z,t)}\ell_{\mathsf{L}}(t) &= -\frac{1}{2}q_Z(z,t)\bar{A}^\top(z,t)D^{-1}\bar{A}(z,t)-\sum_{i=1}^{N}q_Z(z,t_i)\frac{1}{2}\Sigma_x^{-1}\delta(t-t_i).\end{aligned}$$

The solutions to these impulsive differential equations are found as as above, where the jump conditions read

$$\begin{aligned}\lambda(z,t_i) &= q_Z(z,t_i)\Sigma_x^{-1}(x(t_i)-\mu(z,t_i))+\lambda(z,t_i^-)\\ \Psi(z,t_i) &= -q_Z(z,t_i)\frac{1}{2}\Sigma_x^{-1}+\Psi(z,t_i^-).\end{aligned}$$

To solve all derived ODEs, we employ established numerical solvers with adaptive step-size [25, 112].

4.4.3 *Optimizing the variational parameters*

Thirdly, the variational parameters can be optimized iteratively via gradient methods, as is standard for optimal control problems [111]. Here, a simple gradient ascent scheme is used: for each $u(z,t)\in\{A_q(z,t),b_q(z,t),\tilde{\Lambda}(z,z',t)\}$, update

$$u_{\text{new}}(z,t) = u(z,t)+\kappa(z,t)\cdot\partial_{u(z,t)}\ell(t).\tag{4.4.24}$$

To determine the step size $\kappa(z, t)$, we utilize a heuristic back-tracking line-search algorithm [127]: first, we set

$$\kappa(z, t) = \kappa_i q_Z(z, t),$$

that is, we regularize the step size with the respective mode marginal probability. We then choose $\kappa_i = \gamma^i$ with some $\gamma \in (0, 1)$[1]. If $\mathsf{L}[u_{\text{new}}] \geq \mathsf{L}[u]$, we accept the update $u(t) \leftarrow u_{\text{new}}(t)$. Otherwise, we iterate and re-compute using the new step size κ_{i+1}.

The full optimization problem Eq. (4.3.18) requires the Lagrange multiplier ODEs (4.4.20) and (4.4.23) and the constraint Eqs. (4.3.15) and (4.3.17) to be solved jointly as a boundary-value problem with terminal conditions on the multiplier functions $\nu(\cdot, T), \lambda(\cdot, T), \Psi(\cdot, T) = 0$ and initial conditions on the distribution parameters; the variational parameters have to be optimized simultaneously [126]. Boundary-value problems of this type are non-trivial to solve, as both initial and terminal conditions have to be met jointly. A standard approach to this problem is a forward-backward sweeping algorithm, where the solution is again found iteratively [111, 128]: the initial value ODEs for q_Z, μ, Σ and the terminal value ODEs for ν, λ, Ψ are solved alternately until stationarity in the solutions - and that is, stationarity of the ELBO - is reached.

The full algorithm to determine the optimal variational distributions finally is as follows: first solve the Lagrange multiplier ODEs Eqs. (4.4.20) and (4.4.23) backward in time, starting from the terminal conditions $\nu, \lambda, \Psi = 0$. Next, update the variational parameters acting on the constraints, $A_q, b_q, \tilde{\Lambda}, \phi$. Then, solve the constraint equations (4.3.15) and (4.3.17) forward in time, starting from initial conditions $\phi = \{q_Z(z, 0) = q_Z^0(z), \mu(z, 0) = \mu^0(z), \Sigma(z, 0) = \Sigma^0(z)\}$. Finally, repeat until convergence.

4.4.4 *Learning the model parameters*

To learn the parameters of the original model - the prior transition rate matrix Λ, the dispersion D, the prior initial conditions, the parameters of the drift function $f(y, z, t)$ as well as the parameters of the observation likelihood - we employ a variational expectation maximization (VEM) scheme, interleaving the variational optimization described above with the optimization of point estimates for these parameters [10].

After converging onto variational distributions $q_Z(z, t)$ and $q_Y^z(y, t)$, we perform gradient ascent with respect to the model parameters on the ELBO L. The explicit gradients are provided in Appendix C.3. For an in-depth discussion on parameter optimization and more advanced options, see, e.g., [129]. The complete optimization scheme is summarized in Fig. 4.1.

4.5 RESULTS

In the following, the method is evaluated on ground-truth and molecular dynamics (MD) benchmark data. For experimental details such as initializations and hyperparameter settings, see Appendix C.4. Notice that for all shown examples, we empirically observed fast convergence in fewer than 100 iterations of the variational algorithm Fig. 4.1.

[1] Here, $\gamma = 0.5$ is used.

```
input: observation data $\{t_i, x_i\}_{i=1,\ldots,N}$
Initialize $q_Z$, $\mu$, $\Sigma$, $A$, $b$, $\tilde{\Lambda}$, $\Theta$
while $\mathsf{L}$ not converged do
    while $\mathsf{L}$ not converged do
        Compute multiplier functions $\lambda$, $\Psi$, $\nu$ via Eqs. (4.4.20) and (4.4.23)
        Update variational parameters $A_q$, $b_q$, $\tilde{\Lambda}$ via Eq. (4.4.24)
        Compute variational factors $\mu$, $\Sigma$, $q_Z$ via Eqs. (4.3.15) and (4.3.17)
        Update lower bound $\mathsf{L}$
    end
    Update prior parameters via gradient ascent
    Update lower bound $\mathsf{L}$
end
```

FIGURE 4.1: VI algorithm for MJP-switching stochastic differential equations

4.5.1 *Model validation on ground-truth data*

To gauge the accuracy and the limits of the proposed approximation, it will in the following be extensively evaluated against the ground-truth PDE solution as well as against the results obtained using the previously presented Gibbs sampling approach. To that end, it is applied to synthetic data similar to that of Section 3.4.1.

4.5.1.1 *1D system*

We generate synthetic data as in Section 3.4.1: we choose linear, time-homogeneous drift functions $f(y, z, t) = A(z)y(t) + b(z)$ as well as time-homogeneous rates $\Lambda(z, z', t) = \Lambda(z, z')$.

As shown in Fig. 4.2, the inferred posterior distributions $q_Y(y, t) = \sum_{z \in \mathcal{Z}} q_Z(z, t) q_Y^z(y, t)$ and $q_Z(z, t)$ both faithfully reconstruct the respective latent ground-truth trajectories. Accordingly, the obtained parameter estimates exhibit a high accuracy. The MCMC posterior histograms cover the ground-truth still better, which is not surprising as, in that approach, no additional approximation on the process level is made. Notably, the runtime for the VI method was of the order of an hour compared to roughly 24 hours to generate $N = 100000$ samples with the MCMC method.

LIMITATIONS OF THE APPROXIMATION Importantly, Fig. 4.2 also demonstrates the impact of the proposed approximation on reconstruction quality: in regions around mode transitions, one can observe artifacts from the variational approximation as a mixture of GPs. It is apparent that the mixture of GPs results in jumps in the marginal $q_Y(y, t)$ (arrow 1 in Fig. 4.2), while the SSDE is guaranteed to yield continuous paths $y_{[0,T]}$. While at this time point, the approximation still correctly recovers the Z-switch, arrow 2 in Fig. 4.2 indicates a point where the variational posterior $q_Z(z, t)$ misses the transition (caused by three observations that are

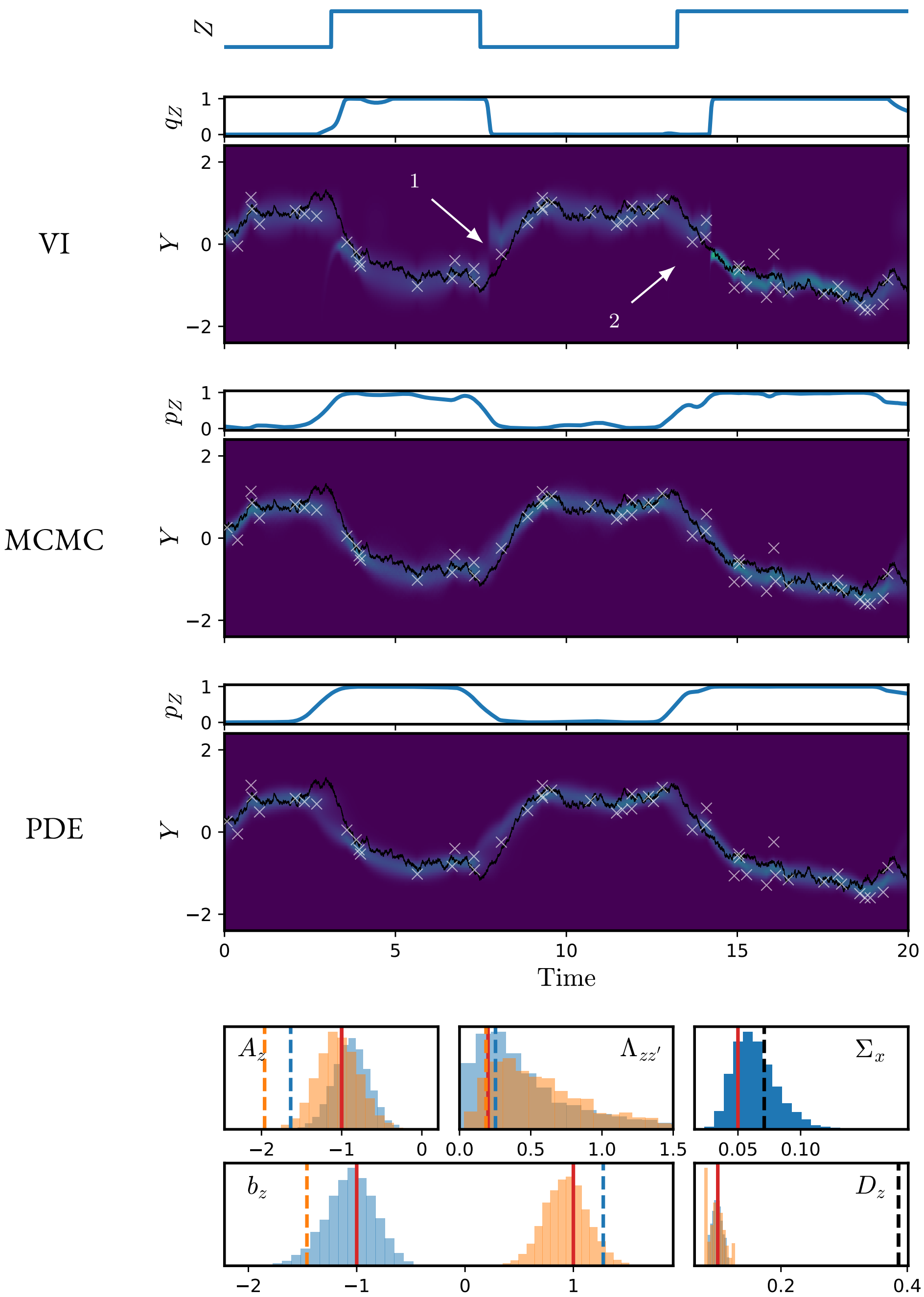


FIGURE 4.2: Model validation of the VI method on 1D synthetic data. Top: ground-truth switching trajectory $z_{[0,T]}$. Following three rows: results for both the discrete and continuous components obtained via the VI approximation (first row), the MCMC framework of Chapter 3 (second row), as well as the exact solution given by the posterior HME (3.2.20) (third row). White arrows mark positions where the approximation yields artifacts, see main text. Bottom: obtained parameters. Vertical red lines mark the ground-truth values; histograms show the MCMC posterior samples; and dashed lines correspond to the point estimates obtained from the VI method (color coding indicates the mode z where applicable).

mis-classified). These inaccuracies are confined to transition regions; since the relaxation onto the mode fixed points is fast compared to the mode remain times, the transition regions are short, yielding a high overall approximation quality. These two transition periods illustrate what was

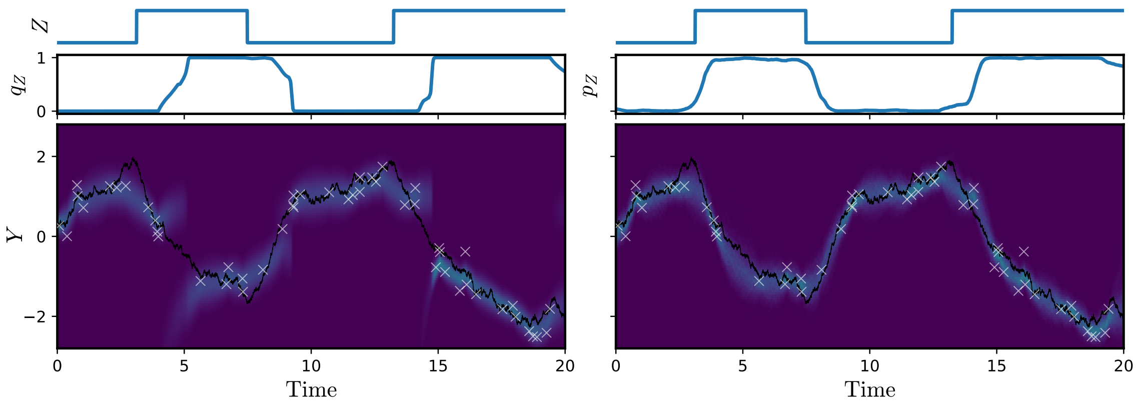


FIGURE 4.3: Inaccuracies of the VI approach in transition regions. Left: VI results. Right: MCMC results. Top row: ground-truth switching trajectory. Middle row: reconstructions of the switching trajectory. Bottom row: Y-space marginals.

discussed above: the variational approximation will be accurate when the system under study has reached a local steady state and it will be inaccurate in the transition periods between individual stable modes. It hence is particularly suited for metastable systems.

To emphasize this point, we repeat the same experiment with slower relaxation constants: $A(z) = -0.5\,\forall z$ (instead of $A(z) = -1$ in the preceding example). As can be seen in Fig. 4.3, the discussed issues become more apparent in the transition regions. Note further that the q_Z-reconstruction appears to be "greedy": the transitions between individual modes are not as smooth as both in the MCMC and the exact PDE solutions. Rather, the method tends to select one of the available modes at each time point.

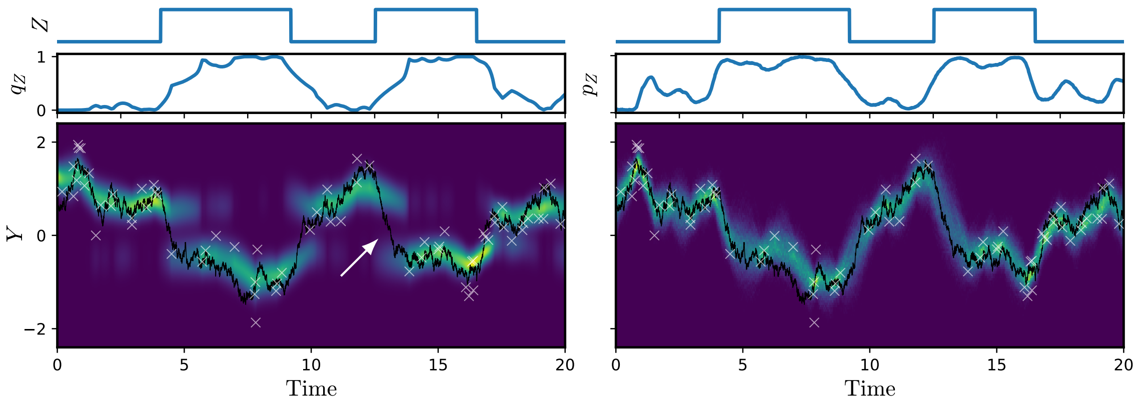


FIGURE 4.4: Performance of the VI approach under higher ambiguity. Left: VI results. Right: MCMC results. White arrow: transition region with high uncertainty.

However, in cases with higher ambiguity, the method is also able to capture the uncertainty in the switching process: we repeat again the experiment with larger ground-truth dispersion and observation covariance. We hence now face a situation with increased ambiguity. The method succeeds in capturing this ambiguity: the reconstruction of the switching sequence $z_{[0,T]}$ is accurate, while showing gradual transitions in regions where no observations are present. In

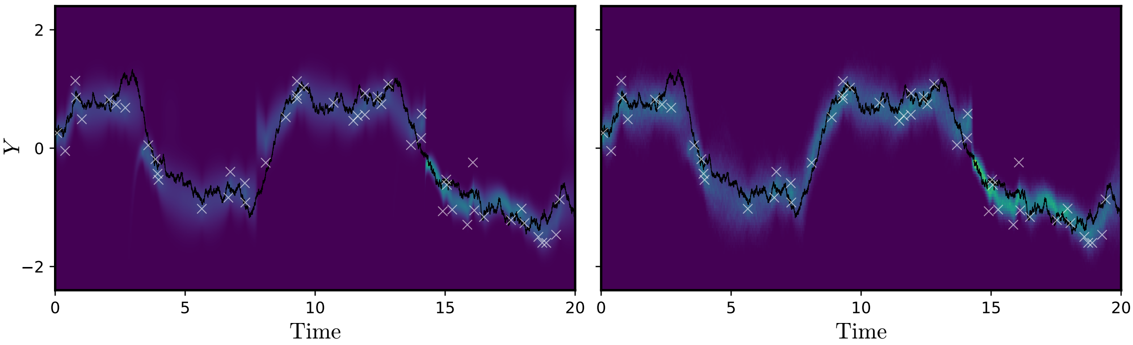


FIGURE 4.5: Sampling from the variational measure. Left: sampling from the fit variational measure as defined in Eq. (4.3.10). Right: sampling from a true SSDE parameterized via the fit variational measure.

particular, note here the long transition period starting at $t = 12.5$ (indicated by the white arrow in Fig. 4.4), where the variational approximation gradually shifts from one GP to the other.

Q-PARAMETERIZED MJP-SSDE SAMPLING An interesting result is obtained by reinterpreting the variational model. Pursuing a generative modeling approach, we can sample full trajectories from the variational posterior; this would recover the marginal density shown in Fig. 4.2 empirically. Specifically, this would entail the discontinuities at Z-transitions. To get rid of these discontinuities, we can - *after* having fit the model - revert back to an SSDE model for Q instead of the mixture of GPs: we can sample from an SSDE which is piece-wise parameterized as the fit variational model. This yields continuous trajectories $y_{[0,T]}$ that tightly follow the ground-truth, as shown in Fig. 4.5.

STRUCTURED MEAN-FIELD VS. MEAN-FIELD We now compare the presented structured mean-field with the classic mean-field approach. To this end, we repeat the experiment shown in Fig. 4.2 three times for each approach with a different number of parameters kept fixed: in one setting, we fix both the mode dynamics A and the observation covariance Σ_x to the true values; in another, we fix only Σ_x; and in the third, we learn all parameters. On one end of the spectrum, the classic mean-field approach returns the most accurate results when both A and Σ_x are fixed. This is unsurprising, as the true latent diffusion, conditioned on the switching process Z, corresponds to a single GP, as does the variational path measure. On the other end of the spectrum, however, the structured mean-field approach yields more accurate reconstructions if all parameters are learned. This, too, is unsurprising, as the proposed approach provides more structure by definition, which can make up for some uncertainty in the parameters.

4.5.1.2 *2D system*

Lastly, we apply the method to the same 2D swirl data as presented in Section 3.4.1. Because the problem itself is more complex, we fix the observation covariance Σ_x to the ground-truth value. As shown in Fig. 4.7, the mode and state reconstructions accurately recover the true paths. Accordingly, also the underlying mode dynamics are correctly learned, exhibiting the counter-rotating behavior of the ground-truth model. Despite the fixed observation covariance, the classic mean-field approach fails to recover the latent switching behavior and accordingly

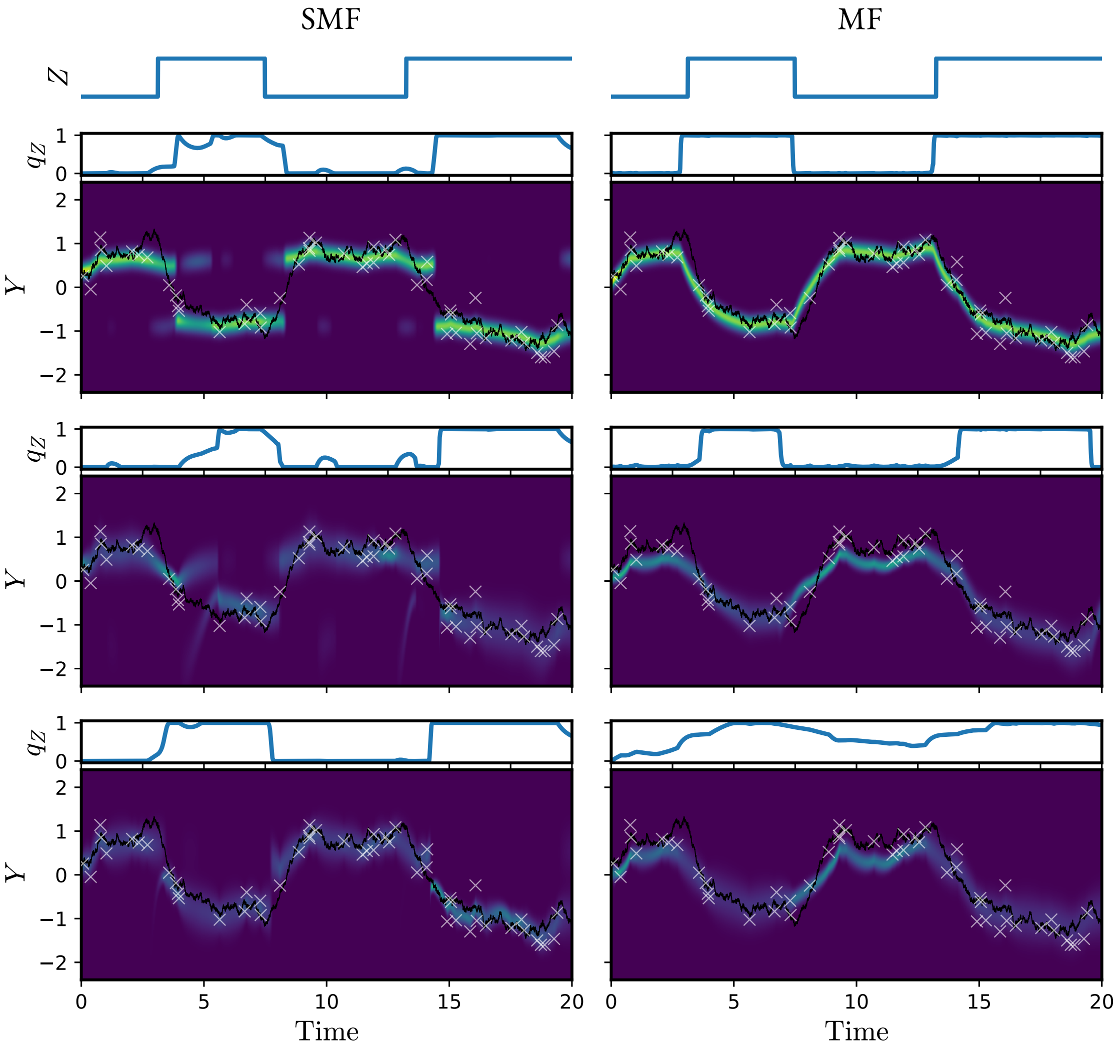


FIGURE 4.6: Comparison of structured mean-field and mean-field approximations. Left column: structured mean-field approximation as presented in this chapter. Right column: mean-field approximation. Top row: fixed slopes A and observation covariances Σ_x. Middle row: fixed observation covariances Σ_x. Bottom row: all parameters learned.

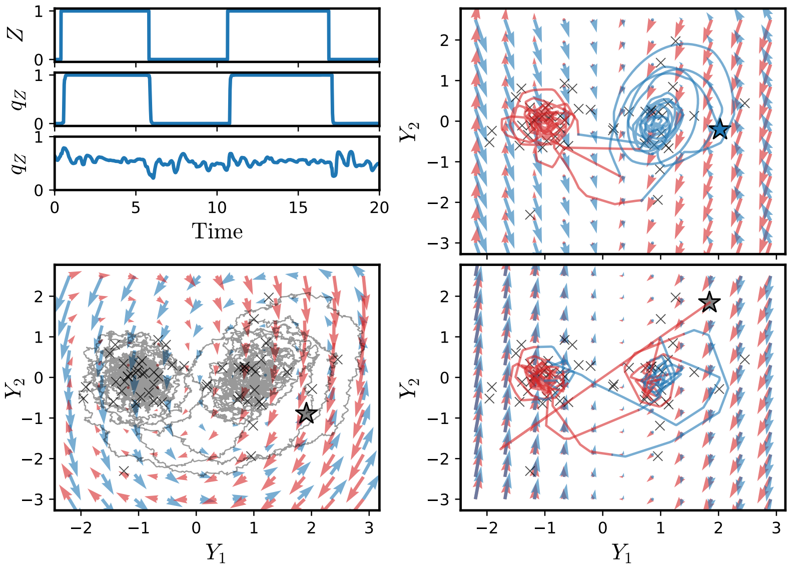


FIGURE 4.7: Model validation of the VI method on 2D synthetic data. Top left: ground-truth switching trajectory (top), structured mean-field (middle) and classic mean-field (bottom) reconstructions. Bottom left: true state trajectory $y_{[0,T]}$, observations (crosses) and true mode dynamics $f(y, z)$ (arrows) in the phase plane. Note that only 20% of observations are shown to reduce clutter. Top right: structured mean-field reconstructed maximum a posteriori (MAP) path $y_{[0,T]}^{\mathrm{MAP}}$ (coloring according to $z_{[0,T]}^{\mathrm{MAP}}$) and reconstructed mode dynamics (arrows). Bottom right: classic mean-field reconstruction (coloring according to $z_{[0,T]}^{\mathrm{MAP}}$).

is not able to discern the qualitatively different dynamics of the two modes. To adequately reproduce the latent dynamics, the time-dependent parameters of the single approximate GP would need to qualitatively change rapidly at a transition (e.g., the sign structure of A), which cannot be achieved at least with a simple gradient learning scheme. Importantly, the runtime advantage of VI increases in higher dimensions: while the MCMC method required around two weeks to compute, the VI approach took less than one day.

4.5.2 *Diffusions in multi-well potentials*

In many real-world scenarios, ground-truth discrete modes may not exist, but the continuous dynamics still exhibit "switching" behavior, transitioning between clearly discernible, qualitatively different regimes. In this case, explicit probabilistic modeling of a set of underlying discrete modes can greatly aid interpretability and enable targeted interventions on the system. The phenomenon of metastability enabling the structured mean-field approximation is often met in biological systems, e.g., in the folding dynamics of complex biomolecules [130]. The folding of proteins or RNA is a continuous process evolving in highly complex potential landscapes. Each local minimum in such a landscape corresponds to a different 3D conformation of the respective molecule, where different conformations can differ qualitatively in their functionality. The relaxation time scales within a local minimum are typically much shorter than between separate minima. To demonstrate the proposed method's capacity to yield sensible representations of

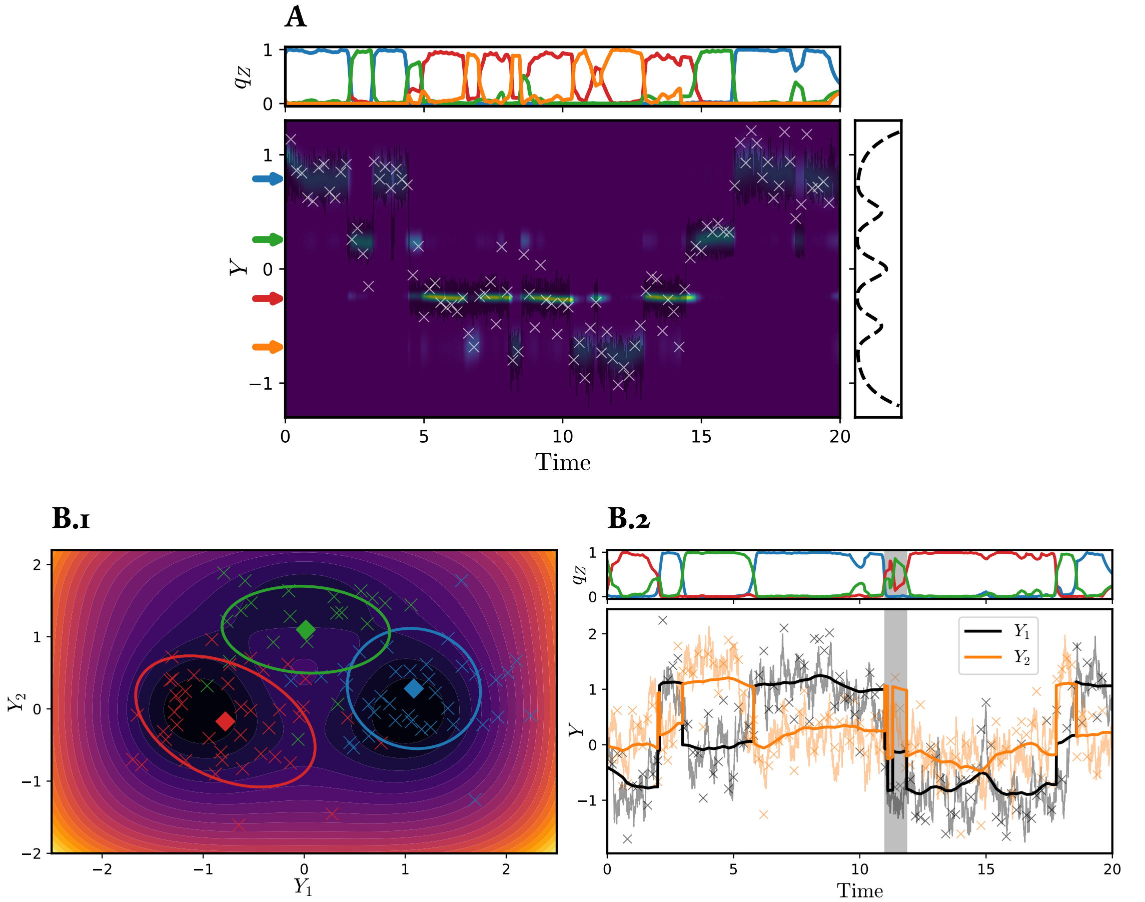


FIGURE 4.8: VI model validation on multi-well diffusion data. **A**: Diffusion in a 1D, four-well potential. Top: inferred marginals $q_Z(z,t)$. Bottom: ground-truth trajectory $y_{[0,T]}$ (black line) and observations (crosses) with the inferred $q_Y(y,t)$ and learned steady states $A^{-1}(z)b(z)$ (arrows). Right: 1D potential landscape. **B**: Diffusion in a 2D, three-well potential. **B.1**: Potential landscape with the inferred steady states (diamonds) and dispersions $D(z)$ (ellipses, 3σ-region) with observations (crosses); colors according to $z^{\mathrm{MAP}}(t_i)$ for each observation $x(t_i)$. Darker colors indicate lower potential values. **B.2**: Top: inferred marginals $q_Z(z,t)$. Bottom: components of $y^{\mathrm{MAP}}(t)$ (thick lines), ground-truth path (thin lines) and observations (crosses). Shaded region: transition region with high ambiguity.

distinct dynamic regimes, we apply it to data generated from latent diffusion processes driven by 1D and 2D benchmark potentials widely used in computational biology [131–136].

The one-dimensional potential, $y \in \mathbb{R}$, exhibits four minima and is given as

$$V(y) = 4\left(y^8 + 3\exp\left\{-80y^2\right\} + 2.5\exp\left\{-80(y-0.5)^2\right\} + 2.5\exp\left\{-80(y+0.5)^2\right\}\right). \tag{4.5.25}$$

The two-dimensional example, $y \in \mathbb{R}^2$, exhibits three minima and reads

$$V(y) = 3\exp\left\{-y_1^2 - \left(y_2 - \frac{1}{3}\right)^2\right\} - 3\exp\left\{-y_1^2 - \left(y_2 - \frac{5}{3}\right)^2\right\} - 5\exp\left\{-(y_1-1)^2 - y_2^2\right\} - 5\exp\left\{-(y_1+1)^2 - y_2^2\right\} + \frac{1}{5}y^4 + \frac{1}{5}\left(y_2 - \frac{1}{3}\right)^4. \tag{4.5.26}$$

In both cases, we fix the observation covariance. Synthetic data are generated from these potentials via the Euler-Maruyama approximation to the SDE

$$\mathrm{d}Y(t) = -\partial_Y V(Y(t)) + Q\mathrm{d}W(t).$$

We assume a mode-dependent dispersion, $D = D(z) = Q(z)Q(z)^\top$.

In both cases, the reconstructed discrete trajectories accurately capture the global transitions between distinct potential minima, see Fig. 4.8. Especially the true (continuous) trajectory in the one-dimensional example, cf. Fig. 4.8 A, exhibits a very clear separation of time scales between the inter- and intra-well dynamics, which is aptly reflected in sharp transitions in the mode reconstruction. In the 2D case, on the other hand, the transition periods are longer between the distinct regions, see Fig. 4.8 B.2. In these transition regimes, it is not possible to unambiguously assign the state at a given time t to one of the three minima. The posterior marginals $q_Z(z, t)$ capture this uncertainty, which is also reflected in a high-quality mode-assignment of the observed data points $x_{[0,T]}$ as shown in Fig. 4.8 B.1. All learned parameters are provided in Appendix C.4.

4.6 SUMMARY

In this chapter, a structured mean-field variational inference framework for MJP-SSDE processes was proposed. The approach was derived starting from the true KL divergence between hybrid processes; it resulted in a straightforward, easily interpretable mixture of GPs by neglecting a term in the KL divergence stemming from the switching component.

The framework was evaluated extensively both on ground-truth and benchmark data. It was shown that a tradeoff exists between faster runtimes (compared to the Gibbs sampler presented in Chapter 3) and (in particular) parameter accuracy: while the variational method is able to learn the system parameters fast and sufficiently well to recover the latent dynamics, the MCMC parameter estimates were somewhat more accurate - which is not surprising, since in this case, no further approximations need to be made. The method also performed well on standard benchmark problems, highlighting that the approximation is applicable to metastable systems, which exhibit a separation of time scales between the continuous and discrete relaxation dynamics. This is a criterion met by many biological systems. An outlook and potential avenues for further work are discussed in Chapter 6.

5

NONPARAMETRIC INFERENCE FOR CONFORMATIONAL SWITCHING

In the previous chapters we analyzed systems in which the latent continuous component was of interest. There are settings, however, in which the continuous dynamics do not matter to analysis. Consider the folding problem of molecules, such as RNA or proteins, already alluded to previously: biological cells operate via the production, specialized utilization and degradation of a vast array of such complex molecules. Importantly, the proper production of any given molecule alone does not guarantee its functionality. Whether or not an RNA can be translated into a protein, or a protein can be put to use by the cell, is determined by the relative spatial arrangement of its constituent sub-molecules, that is, its folding structure or *conformation*. Due to the crucial impact of the conformational arrangement, computational structure prediction is a prime goal of molecular biology. In recent years in particular, novel frameworks were proposed for RNA [137] and protein [138] structure prediction that yield unprecedented prediction accuracies; the focus of these structure prediction algorithms however lies on *static* structures. Different conformations can be thought of as minima in a high-dimensional complex energy landscape through which the molecule evolves over time, for instance upon receiving an external input signal such as a binding ligand or by thermodynamic fluctuations [139]. Importantly, the particular dynamics within a given minimum do not matter to conformational analysis, but the transitions between different minima are of central interest as they may change the molecule's functionality.

The challenge is then (i) to identify the energy minima (in particular, their number), and (ii) to model the dynamics between them.

Conformational dynamics may be studied both experimentally and computationally. A prime example of experimental approaches is the analysis of ion channel gating behavior via electrophysiological techniques, such as voltage clamping or lipid bilayer measurements [140–142]. Further techniques include fluorescence quenching [143] or fluorescence resonance energy transfer (FRET) measurements [144]. The dominating tools for computational analyses, on the other hand, are molecular dynamics (MD) simulations [145]. The MD framework aims at the mechanistic simulation of the interaction of all particles constituting the molecule under study, where "particles" may refer to atoms (in classic MD) or to atomic or even molecular clusters in coarse-grained MD [146]. Simulation techniques of this kind are ubiquituously used in molecular biology to investigate protein and RNA folding [147–149], with conformational switching of ion channels being an example of the former [150].

As both electrophysiological data and MD simulation data are typically obtained on a regular time grid with a very small time step Δt, the following chapter adopts a discrete-time approach. A popular discrete-time framework to address the above two challenges are MSMs, which yield a description of the *continuous* system dynamics (e.g., 3D atom coordinates) in terms of comparatively long-lived, metastable *discrete* states corresponding to distinct, stable structural conformations [131, 132, 135, 151]. As will be discussed in detail in Section 5.1, this framework however suffers from two important drawbacks, namely, (i) the need to manually preprocess the data, and (ii) the need to manually identify the number of metastable states.

In this chapter, these two drawbacks are addressed utilizing nonparametric Bayesian hidden Markov models (HMMs): we model the conformational switching between distinct molecular structures via an HMM by defining a latent discrete-time Markov chain (DTMC) on a countably infinite set of states of which noisy, continuous-valued observations are obtained. Nonparametric Bayesian approaches have gained attention in recent years both in experimental settings such as analyses of ion channel switching [152, 153] or single-particle tracking [154] as well as in MD studies [155]. In all of these cases, inference of the metastable trajectories and the system parameters is carried out using sampling techniques (such as MCMC), which - as discussed in the previous chapters - are known to face scaling issues [156, 157]. Even for the relatively simple problem of one-dimensional ion channel voltage trajectories, they become computationally intractable for longer sequences. To achieve scalability to large data sets, we combine the nonparametric model with a variational inference approach as done, e.g., in [158].

We specify observation models that are appropriate for the use cases laid out above: in the described experimental and computational settings, observations typically are real-valued vectors, $x \in \mathbb{R}^n$, or rotation angles, $x \in [0, 2\pi)^n$. For the angular case, we furthermore propose an approximation enabling computational tractability.

In the following, a coarse overview over MSMs will be given in Section 5.1 to provide more intuition about the problem of conformational switching, along with the shortcomings of the method that we aim to address. Then, the general nonparametric modeling framework will be defined in Section 5.2, followed by a discussion of the specific observation models. The details on the variational inference scheme are subsequently laid out in Section 5.4. Finally, the method will be applied to synthetic ground-truth data as well as benchmark and experimental data.

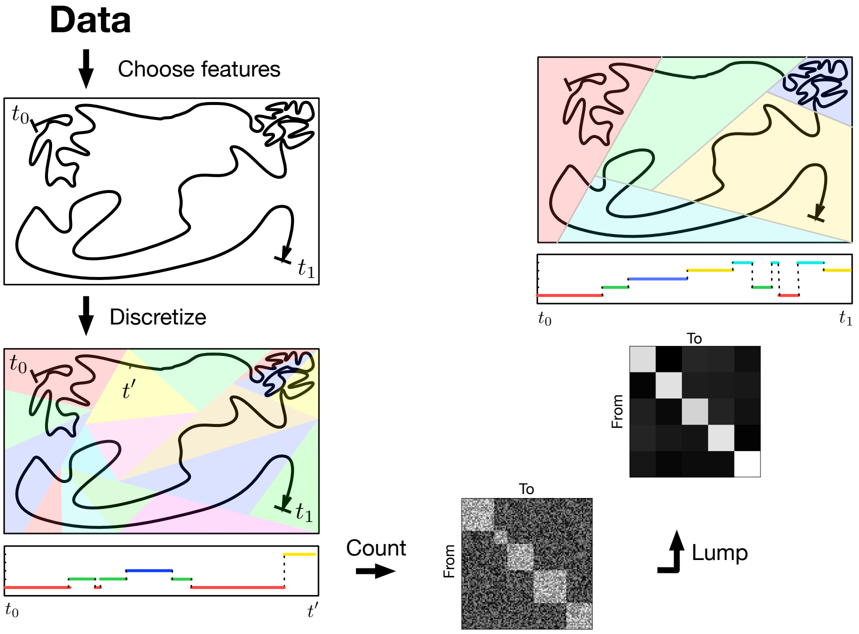


FIGURE 5.1: Sketch of the workflow for classic MSMs. Given some data (e.g. from MD simulations), one chooses a set of suitable features that allow to compactly parameterize the data. Note that for illustrative purposes, only a single trajectory is shown, where in typical applications, many trajectories are analyzed jointly. The state space thus defined is then discretized. This can for instance be done via k-means. In that way, the data is projected onto a discrete trajectory, from which via straightforward counting an empirical transition matrix between the discrete states can be obtained. The resulting "fine-grained" DTMC is then lumped into a "coarse-grained" one consisting of only a few macrostates; the lumping is typically done via spectral methods such as Perron-cluster cluster analysis (PCCA). Finally, this DTMC constitutes the MSM on a discrete state space of only a few, macroscopically relevant and interpretable states.

5.1 MARKOV STATE MODELS

The classic MSM approach approximates the continuous molecule dynamics directly by projecting them to a discretized space with only a handful of states, the dynamics of which are described as a DTMC (cf. Section 2.1.1.1). This achieves two central goals: first, it yields a readily interpretable, parsimonious system description, enabling the (principled) handling of very large MD data sets. Second, it can help to drastically reduce the required computation time for MD simulations: MSM analysis can be applied to heavily parallelized data. Rather than using one single very long MD trajectory, one may hence start many MD simulations in parallel with different initializations and analyze these trajectories jointly. In this way, a conformational transition model can be obtained much faster.

While conceptually straightforward, the process of MSM construction is a multi-step procedure; see Fig. 5.1 for a sketch of the basic workflow. One issue in this process is the direct binning of continuous data into discrete states, which introduces correlations and makes the resulting discrete, projected-data process generally non-Markovian. Raw MD data, on the other hand, are inherently Markovian, as they originate from the integration of SDEs, cf. Section 2.1.2.2. To render the discrete dynamics amenable to MSM analysis - that is, to ensure that the projected-data process can be appropriately modeled by a Markov process - the correlations need to be reduced via temporal thinning by some *lag time* constant τ [131, 159]. Concretely, to construct the fine-grained transition matrix of Fig. 5.1, one does not count transitions between neighboring time points but between time points separated by τ time steps. The selection of this parameter is acknowledged to be a major challenge in practice [160].

The lag time, together with the state-space discretization, determines the reconstruction error between the true system dynamics and their MSM approximation. It is possible to explicitly derive an upper bound to this error, showing that it can be made arbitrarily small by either choosing a finer state-space discretization or increasing the lag time [131]. Accordingly, this however generates a trade-off problem: one has to balance between (i) sufficient sampling of the discrete state space, and (ii) a sufficiently long lag time to render the resulting process Markovian. To gauge the appropriateness of a given τ, tools such as the Chapman-Kolmogorov test have been introduced [131]. However, this test only considers the appropriateness of the lag time, which still allows for a considerable reconstruction error - resulting in an MSM not accurately reproducing the long-time dynamics as detailed in [161].

The other issue of the MSM workflow is the identification of metastable states itself, which is carried out as the last step. Typically, this lumping of a DTMC on a large state space to a DTMC on a much smaller state space is done via spectral methods such as Perron-cluster cluster analysis (PCCA) [162] or PCCA+ [163] (note, however, that other approaches exist, see, e.g., [164, 165]). Generally, the chosen lag time as well as the state-space discretization will affect the lumping results.

Fueled by the successes of MSM modeling in a wide array of use cases [166–173], extensive methodological research has been carried out in recent years [131, 135, 136, 174–176]. The most recent advances also include deep learning extensions [132, 177]. Nevertheless, they share the two conceptual drawbacks detailed above; both problems can however be addressed utilizing nonparametric HMM frameworks, as will be discussed in the following.

5.2 NONPARAMETRIC BAYESIAN MARKOV STATE MODELS

We model the conformational molecule dynamics by utilizing an HMM consisting of two joint stochastic processes $\{Z(t), X(t) : t = 1, \ldots, T\}$. The distinct metastable states are represented by the latent Markov states $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$, the observed data (e.g., experimentally obtained channel voltages or simulated atom positions) by $X(t) \in \mathcal{X} \subseteq \mathbb{R}^n$. The system dynamics are described by the transition matrix of the DTMC, see Section 2.1.1.1. The observation at time point t, $X(t) = x$, depends only on the latent state at the same time, $Z(t) = z$, via some observation density $p(x \mid \theta_z)$.

To obtain a fully Bayesian HMM, the transition matrix as well as the observation parameters are treated as random variables: we define

$$\begin{aligned} \Pi_z &\sim \mathrm{Dir}(\eta_{z,1}, \ldots, \eta_{z,|\mathcal{Z}|}), \\ \Theta_z &\overset{\text{i.i.d.}}{\sim} \mathsf{P}_0, \end{aligned} \tag{5.2.1}$$

for all $z = 1, ..., |\mathcal{Z}|$ with $\eta_{z,z'} > 0$. A realization $\Pi_z = \pi_z$ can hence be interpreted as the z-th row of the transition matrix, with the entry z' denoting the probability to transition from state z to state z':

$$\pi_{zz'} = \mathsf{P}(Z(t+1) = z' \mid Z(t) = z).$$

We furthermore assume that P_0 admits a probability density $p(\theta)$. In the following, we denote with

$$\left\{x^i\right\} := \left\{x^i_{[1,T]} : i = 1, \dots, I\right\}, \qquad \left\{z^i\right\} := \left\{z^i_{[1,T]} : i = 1, \dots, I\right\}$$

the set of all I observed and latent trajectories

$$x^i_{[1,T]} := \left\{x^i(t) : t = 1, \dots, T\right\}, \qquad z^i_{[1,T]} := \left\{z^i(t) : t = 1, \dots, T\right\}.$$

Analogously, $\{\pi_z\} := \{\theta_z : z = 1, \dots, |\mathcal{Z}|\}$ and $\{\theta_z\} := \{\theta_z : z = 1, \dots, |\mathcal{Z}|\}$. With these definitions, the joint model density reads

$$\begin{aligned} p\left(\left\{x^i\right\}, \left\{z^i\right\}, \{\theta_z\}, \{\pi_z\}\right) = \prod_{i=1}^{I} p(z^i_1) p(x^i_1 \mid \theta_{z^i_1}) \prod_{t=2}^{T} p(z^i_t \mid \pi_{z^i_{t-1}}) \\ \cdot\, p(x^i_t \mid \theta_{z^i_t}) \prod_{z=1}^{|\mathcal{Z}|} p(\pi_z) p(\theta_z), \end{aligned} \tag{5.2.2}$$

where we abbreviate $z^i(t) = z^i_t$ for conciseness (and analogously for x).

It is acknowledged that HMMs can be interpreted as a generalization of MSMs [136], abolishing the need to introduce an artificial lag time. This is due to the fact that marginalization over Z generates global dependencies between the observations; hence, the observed marginal process X does not need to be Markovian.

HMMs however still retain the drawback with respect to the analysis of conformational switching that the number of molecule conformations $|\mathcal{Z}|$ needs to be specified upfront, while this number is typically unknown. On the contrary, it is a key quantity of interest that is to be determined from the data. This issue may be addressed by taking a nonparametric modeling approach, which allows for specification of processes on countably infinite state spaces and computation of the respective posteriors. For any finite data set, the posterior state-space size $|\mathcal{Z}|$ will still be finite and can be learned from the observations. In other words, we specify a model for a potentially infinite number of distinct molecular conformations, only a finite number of which will be adopted in any given observed trajectory from simulations or experiments. This model addresses both of the key drawbacks of conventional MSMs: introducing a latent process obviates the need to manually pre-process the data with some artificial lag time, while the nonparametric nature allows one to directly learn the number of conformational states from the data directly rather than having to resort to manual post-hoc analysis.

In order to set up a nonparametric HMM, we accordingly need to construct prior distributions for transition matrices on countably infinite state spaces and for countably infinite observation parameters; put differently, the distributions occurring in Eq. (5.2.1) need to be generalized to $|\mathcal{Z}| \to \infty$. This is achieved by the hierarchical Dirichlet process (HDP) [38]. An HDP-HMM is constructed in two steps utilizing Dirichlet processs (DPs) (see Section 2.2.3): first, we specify a DP via the stick-breaking construction Eq. (2.2.46),

$$\mathsf{P}_1 \sim \mathrm{DP}(\gamma, \mathsf{P}_0) \rightarrow \begin{cases} \mathsf{P}_1(\Theta = \theta) = \sum_{z=1}^{\infty} B_z \delta_{\Theta_z}(\theta), & (5.2.3) \\ B \sim \mathrm{GEM}(\gamma), & (5.2.4) \\ \Theta_z \overset{\text{i.i.d.}}{\sim} \mathsf{P}_0, & (5.2.5) \end{cases}$$

where it is assumed (as above) that P_0 admits a probability density $p(\theta)$. This measure determines a prior over conformations: each z represents a distinct structure, with B_z its probability and Θ_z its associated parameterization.

Second, the stick-breaking measure B serves in turn as the base measure of another, subordinate DP: consider independent random variables

$$\Pi_z \sim \mathrm{DP}\left(\kappa + \xi, \frac{\kappa B + \xi \delta_{\Theta_z}}{\kappa + \xi}\right), \tag{5.2.6}$$

with the *stickiness* parameter $\xi \geq 0$ on which will be elaborated in the next paragraph. As before, this measure can be expressed through

$$\begin{aligned} H_z &\sim \mathrm{GEM}(\kappa + \xi), \\ \Phi_{zz'} &\overset{\text{i.i.d.}}{\sim} \frac{\kappa B + \xi \delta_z}{\kappa + \xi}, \end{aligned}$$

with the point measure at index z, δ_z. This allows us to write

$$\Pi_z(Z = z') = \sum_{j=1}^{\infty} H_{zj} \delta_{\Phi_{zj}, z'}.$$

with the (discrete) Kronecker delta instead of the (continuous) Dirac delta function. This measure defines a process on the index set, $Z \in \mathcal{Z}$. Note, however, that by combination with the support of P_1, it can be re-expressed as a process on the parameter space via

$$\Pi_z(\Theta = \theta) = \sum_{j=1}^{\infty} H_{zj} \delta_{\Theta_{\Phi_{zj}}}(\theta),$$

making clear that all these processes have shared support via the atoms $\{\Theta_1, \Theta_2, ...\}$ of P_1; $\Pi_z(Z = z')$ refers to the same state z' parameterized by $\Theta_{z'}$ for all z. Each Π_z can hence be understood as a distribution over a row of a "countably infinite transition matrix": each atom Θ_z corresponds to one latent state z - that is, one molecular conformation - and parameterizes the respective observation distribution, $p(x \mid Z = z, \{\Theta_1 = \theta_1, \Theta_2 = \theta_2, ...\}) = p(x \mid \theta_z)$. In other words, the two-step HDP-HMM construction (i) defines the molecular conformations via the measure P_1, and (ii) determines their transition dynamics via all Π_z.

The parameter ξ introduces a self-transition bias: it extends the sojourn times within each state, which is why it is commonly denoted as "stickiness". For $\xi = 0$, the conventional HDP-HMM is recovered [38]. The sticky HDP-HMM has been shown to counter-balance the strong sensitivity of the conventional HDP-HMM to within-state variability, resulting in a tendency to introduce

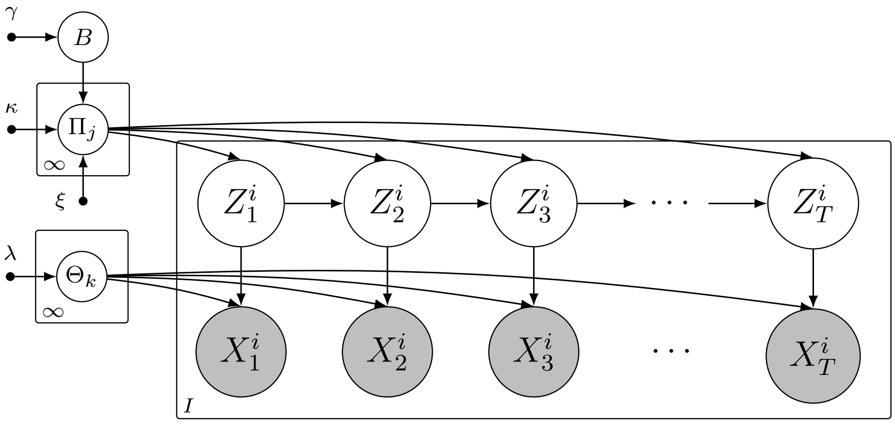


FIGURE 5.2: Probabilistic graphical model of the HDP-HMM. We study I observed trajectories $\{x(t)^i : t = 1, ..., T; i = 1, ..., I\}$ (shaded circles) which we assume to be generated by latent trajectories $\{z(t)^i\}$ (unshaded circles); we abbreviate $z(t) = z_t, x(t) = x_t$. The HDP-HMM defines countably infinite states parametrized by Θ_k, the transitions between which are described via the transition probabilities Π_k. These transition probabilities are drawn from a DP with the base measure B.

redundant states pertaining to the same ground-truth state, see, e.g., [156]. The sticky HDP-HMM consequently matches the setting of interest in the present chapter, as we aim specifically for the analysis of metastable states which may potentially exhibit a high level of intra-state variability in the respective observational data. Note as an aside that with respect to classic MSMs, the stickiness parameter can be interpreted as a bias towards longer time scales that is to be set by the experimenter. Importantly, the present approach does not discard any information, but retains all available data.

In summary, the sticky HDP-HMM is defined as follows, see Fig. 5.2: let $\gamma > 0, \kappa > 0, \xi \geq 0$ and P_0 a probability measure. The tuple $(\{X(t), Z(t) : t = 1, \ldots T\}, \ \{\Theta_z\}, \{\Pi_z\}, B)$ is distributed according to a sticky HDP-HMM,

$$\{X(t), Z(t) : t = 1, \ldots, T\}, \{\Theta_z\}, \{\Pi_z\}, B \sim \text{HDP-HMM}\ (\gamma, \kappa, \xi, \mathsf{P}_0), \qquad (5.2.7)$$

if B, $\{\Theta_z\}$ and $\{\Pi_z\}$ are distributed according to Eqs. (5.2.4), (5.2.5) and (5.2.6) and

$$\begin{aligned} Z(t) \mid Z(t-1) = z, \Pi_z = \pi_z &\sim \text{Cat}(\pi_z), \\ X(t) \mid Z(t) = z, \Theta_z = \theta_z &\sim p(\,\cdot\mid \theta_z). \end{aligned}$$

Note that to avoid clutter, this definition is formulated only for a single sequence. It is the nonparametric analogue to the finite Bayesian HMM; importantly, while we can conceptually formulate a joint model density similar to Eq. (5.2.2), we cannot evaluate it due to the occurring infinite products such as $\prod_{z=1}^{\infty} p(\theta_z)$. It will be seen in Section 5.4 that utilizing a VI approach, this does not hinder our ability to draw inferences about the posterior process.

5.3 OBSERVATION MODELS

To complete the specification of the HDP-HMM, it remains to set up appropriate observation distributions $p(x \mid \theta_z)$ as well as the corresponding priors $p(\theta_z)$.

REAL-VALUED DATA A versatile model for coordinate data $\mathcal{X} \subseteq \mathbb{R}^n$ as often obtained via MD - such as 3D atom positions - as well as in electrophysiological experiments is the multivariate Gaussian used also in the previous chapters,

$$X(t) \mid Z(t) = z \sim \mathcal{N}(\mu_z, \Sigma_z), \tag{5.3.8}$$

with mean $\mu_z \in \mathbb{R}^n$ and covariance matrix $\Sigma_z \in \mathbb{R}^{n \times n}$ for all z. Both MD simulation data as well as experimental voltage trajectory data can be covered with this conventional choice [152, 153, 155].

As discussed in Chapter 3, the respective conjugate prior is the NIW distribution

$$\mu, \Sigma \sim \text{NIW}(\mu_0, \rho, \Psi, \nu) \Leftrightarrow \begin{cases} \mu \sim \mathcal{N}(\mu_0, \Sigma/\rho), \\ \Sigma \sim \text{IW}(\Psi, \nu). \end{cases} \tag{5.3.9}$$

We combine the likelihood Eq. (5.3.8) with the conjugate prior Eq. (5.3.9) to specify the HDP-HMM for real-valued data.

ANGULAR DATA Another way of specifying the spatial arrangement of complex molecules - which is widely used, particularly in MD - are the *dihedral angles* between adjacent atom or molecule planes [145]. The data are then constrained to the unit circle, $\mathcal{X} \subseteq [0, 2\pi)^n$. Appropriate observation models for these spaces are von Mises-type distributions [178, 179]. As it is common to characterize amino acid chains such as proteins by sets of*pairs* of angles (ϕ_i, ψ_i), we will focus on the two-dimensional case and note that the theory readily extends to longer chains. Here, the cosine-variant of the bivariate von Mises (BvM) distribution is utilized; this is an established choice in the context of protein modeling [180]. For $x = (\phi, \psi) \in [0, 2\pi)^2$,

$$\begin{aligned} & X(t) \mid Z(t) = z \sim \text{BvM}(\zeta_z, \nu_z, \lambda_{z,1}, \lambda_{z,2}, \lambda_{z,3}) \\ \Leftrightarrow \quad & p(\phi, \psi \mid \zeta_z, \nu_z, \lambda_{z,1}, \lambda_{z,2}, \lambda_{z,3}) = c(\lambda_{z,1}, \lambda_{z,2}, \lambda_{z,3}) \exp\{\lambda_{z,1}\cos(\phi - \zeta_z) \\ & \quad + \lambda_{z,2}\cos(\psi - \nu_z) - \lambda_{z,3}\cos(\phi - \zeta_z - \psi + \nu_z)\}, \end{aligned} \tag{5.3.10}$$

where

$$c(\lambda_1, \lambda_2, \lambda_3) = \frac{1}{(2\pi)^2}\left[I_0(\lambda_1)I_0(\lambda_2)I_0(\lambda_3) + 2\sum_{k=1}^{\infty} I_k(\lambda_1)I_k(\lambda_2)I_k(\lambda_3)\right]^{-1} \tag{5.3.11}$$

and I_i is the modified Bessel function of the first kind and order i. The location parameters ζ and ν control the position of the modes of the distribution, as can be seen from the trigonometric terms in Eq. (5.3.10); the parameters $\lambda_1, \lambda_2, \lambda_3$ specify the spatial correlations. Note that marginalizing over ϕ and setting $\lambda_1 = \lambda_2 = 0$ recovers the conventional one-dimensional von Mises (vM) distribution,

$$p(\psi \mid \nu, \lambda_2) = \frac{\exp\{\lambda_2 \cos(\psi - \nu)\}}{2\pi I_0(\lambda_2)}. \tag{5.3.12}$$

Analytical expressions for a conjugate prior do exist for the bivariate von Mises distribution [178]. However, the (infinite) sum of Bessel functions in Eq. (5.3.10) renders the normalizer c

intractable in a Bayesian setting, as one needs to compute expectations with respect to the λ-parameters. Also, this distribution does not need to be unimodal: there exist intricate conditions on the relation of the concentration parameters $\lambda_1, \lambda_2, \lambda_3$ to achieve unimodality [179]. It is however known that for high concentration values in specific regimes, the bivariate von Mises distribution is well approximated by a bivariate normal distribution [178]. This is unsurprising because von Mises-type distributions and Gaussian distributions are tightly linked. In fact, the former can be constructed from the latter [181]. Focusing on systems exhibiting distinct, separable metastable states, we expect peaked, highly concentrated angular distributions, agreeing with the requirements of this approximation. To ensure tractability and interpretability, we hence make use of this and approximate

$$\mathrm{BvM}(\zeta, \nu, \lambda_1, \lambda_2, \lambda_3) \approx \mathcal{N}(\mu, \Sigma). \tag{5.3.13}$$

The mode position of the BvM roughly corresponds to the mean vector μ and the covariance depends on the λ-parameters. The precise analytical expressions for these dependencies are involved and not relevant to our approximation - for an in-depth analysis, the interested reader is referred to [178, 179].

Utilizing Eq. (5.3.13), we can employ the same conjugate prior as before, $\mathrm{NIW}(\mu_0, \rho, \Psi, \nu)$. We then deal with the periodicity by projecting the data into an interval $[-\pi, +\pi]$ around the mean μ_0:

$$x \leftarrow x - 2\pi \cdot \mathrm{sgn}(x - \mu_0). \tag{5.3.14}$$

Note that this necessarily leads to an underestimation of the covariance because the data is treated as if it were produced by a normal distribution whereas in reality, it is generated by a von Mises distribution: data outside of $[0, 2\pi)$ do not occur. For two reasons, this is acceptable in the present setting: first, as detailed above, we assume the data to be peaked for the approximation to hold. If this assumption is valid, the probability mass outside the interval $[-\pi, +\pi]$ is negligible. An illustration of this will be presented in the next section, see Fig. 5.6. Second, as we pursue an approximate approach to inference, the obtained results in any case have to be interpreted and judged for accuracy. The gained tractability may greatly increase the practical utility of the framework, as it is otherwise also customary to resort to 3D real-valued coordinates to avoid mathematical complexity, disregarding crucial structural information about the biological problem [155]. In the following, we refer to Eq. (5.3.13) as the *approximate von Mises* model.

5.4 VARIATIONAL INFERENCE OF CONFORMATIONAL STATES

Taking a Bayesian approach, we are interested in the posterior distribution

$$\left\{Z^i\right\}, \left\{\Theta_z\right\}, \left\{\Pi_z\right\}, B \mid \left\{x^i\right\}. \tag{5.4.15}$$

This distribution cannot be evaluated analytically in closed form. In principle, the conjugacy property of the DP (see Section 2.2.3) allows for the utilization of sampling techniques to obtain the posterior empirically [156, 157]. However, the typically large data sets from simulations or long-duration experiments (see, e.g., the discussion in [152]) render this approach computationally infeasible, as every draw from the posterior requires one full pass through the data.

To alleviate these computational issues, we utilize a mean-field variational inference approach, cf. Section 2.2.4.2 and Chapter 4. Since the HDP-HMM specifies distributions with countably infinite support, VI in this case requires an additional variational parameter to be able to actually instantiate the variational distributions. Proceeding as in the previous section, the VI problem is first formulated for the finite Bayesian HMM and then generalized to the infinite case.

VARIATIONAL INFERENCE FOR THE FINITE HMM We aim to identify the variational measure Q minimizing the KL divergence to the true posterior measure P_x,

$$\arg\min_{\mathsf{Q}} \mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}_x], \tag{5.4.16}$$

where P_x clearly admits for a density $p\left(\left\{z^i\right\}, \{\theta\}, \{\pi\} \mid \left\{x^i\right\}\right)$ as can be seen from Eq. (5.2.2). Specifying Q similarly via

$$q\left(\left\{z^i\right\}, \{\theta_z\}, \{\pi_z\}\right) \tag{5.4.17}$$

allows to express the VI problem via p and q directly,

$$\arg\min_{q} \mathsf{D}_{\mathrm{KL}}[q \mid\mid p], \tag{5.4.18}$$

or, equivalently (see Section 2.2.4.2), as

$$\arg\max_{q} \mathsf{L} = \arg\max_{q} \left\{ \mathsf{E}\left[\ln \frac{p\left(\left\{x^i\right\}, \left\{z^i\right\}, \{\theta_z\}, \{\pi_z\}\right)}{q\left(\left\{z^i\right\}, \{\theta_z\}, \{\pi_z\}\right)}\right]\right\} \tag{5.4.19}$$

with the ELBO L, sparing one to having to evaluate the computationally intractable log-evidence.

Employing a mean-field assumption,

$$q\left(\left\{z^i\right\}, \{\theta_z\}, \{\pi_z\}\right) = \prod_{i=1}^{I} q\left(z^i\right) \prod_{z=1}^{|\mathcal{Z}|} q\left(\theta_z\right) q\left(\pi_z\right), \tag{5.4.20}$$

enables an iterative coordinate-wise ascent optimization procedure (cf. Section 2.2.4.2). One variational factor of Eq. (5.4.20) is optimized at a time while keeping all others fixed, yielding the generic distribution update for any variational factor $\alpha \in \left\{\left\{z^i\right\}, \{\pi_z\}, \{\theta_z\}\right\}$:

$$q(\alpha) \propto \exp\left\{\mathsf{E}_{q\setminus\alpha}\left[\ln p\left(\left\{x^i\right\}, \left\{z^i\right\}, \{\pi_z\}, \{\theta_z\}\right)\right]\right\}, \tag{5.4.21}$$

where $\mathsf{E}_{q\setminus\alpha}$ denotes the expectation with respect to all variational distributions except $q(\alpha)$. Note that the ELBO is not convex with respect to all variational distributions jointly [50], while it is convex with respect to any factor individually [9]. This coordinate-wise ascent algorithm hence converges to a local optimum.

VARIATIONAL INFERENCE FOR THE INFINITE HMM To generalize the above to countably infinite state spaces, conceptually only the stick-breaking measure $B \sim \mathrm{GEM}(\gamma)$ needs to be added. It is however clear by inspection of Eq. (5.4.20) that it is necessary to truncate the number of variational states to some maximum K in order to be able to instantiate the analogous mean-field variational ansatz

$$q\left(\left\{z^i\right\},\{\theta\},\{\pi\},\beta\right)=\prod_{i=1}^{I} q\left(z^i\right)\prod_{z=1}^{K} q\left(\theta_z\right)q\left(\pi_z\right)q(\beta). \tag{5.4.22}$$

In principle, this truncation level could be set to the number of data points; in practice, it is convenient for computational reasons to choose some number which is large compared to the expected number of HMM states [156, 182]. This only affects the variational distributions while the original model Eq. (5.2.7) remains unchanged [183]. Due to the defining property of the DP, it however allows to also evaluate the original model density. Constraining the variational posterior to a maximum of K states induces a partition on the base measure space of P_0: recalling the defining property of the DP, Eq. (2.2.43), one can hence write the prior measures Eq. (5.2.6) as finite Dirichlet distributions with $K+1$ dimensions, corresponding to K states and the "rest" of the state space (where no transitions are observed):

$$\Pi_z \mid B=\beta \sim \mathrm{Dir}(\kappa\beta_1+\delta_{z,1},...,\kappa\beta_{K+1}+\delta_{z,K+1}). \tag{5.4.23}$$

The full model can be formally expressed as

$$p\left(\{x^i\},\{z^i\},\{\pi_z\},\{\theta_z\},\beta\right)=\prod_{i=1}^{I} p(z_i^1)p(x_1^i \mid \theta_{z_1^i})\prod_{t=2}^{T} p(z_t^i \mid \pi_{z_{t-1}^i})p(x_t^i \mid \theta_{z_t^i})$$
$$\prod_{z=1}^{\infty} p(\pi_z \mid \beta)p(\theta_z)p(\beta). \tag{5.4.24}$$

This expression contains an infinite product, preventing us from explicit evaluation. Combined with the truncation, this does no harm, however; we choose the *direct assignment* truncation method, setting $q(z,t)=0$ for any $z>K$ [182]. With this assumption in place, the variational problem Eq. (5.4.19) is still well defined, as any parameters θ_z, π_z for $z>K$ do not contribute: all expectations within the KL divergence evaluate to zero.

While other truncation schemes also exist [53, 184], direct assignment is particularly appealing due to its simplicity. The thus-defined variational model can, but does not need to, utilize all clusters up to K [184]. Importantly, this allows for straightforward debugging, as it is directly apparent whether all K states are occupied: if $q(z,t)>0$ for all states z, one might incur a non-negligible truncation error, as - intuitively speaking - more states might be needed to explain the data, and a double-check with increased K is necessary. If, on the other hand, $q(z,t)=0$ for some states, the variational approximation is expressive enough and will not result in a significant truncation error. Note furthermore that the direct assignment truncation can be utilized for automated search algorithms over the truncation depths [185]. In the following, the resulting updates (5.4.21) will be presented.

LATENT STATE SEQUENCE The marginal probabilities of metastable states, $q(z,t)$, can be computed by a forward-backward message-passing algorithm similar to the forward-backward algorithm for exact HMM inference presented in Section 2.2.1.1 [186]. To reduce clutter, the sequence index i is omitted in the following. The forward messages $\alpha(z,t)$ and the backward messages $\beta(z,t)$ are computed as

$$\alpha(z,t) = \exp\left\{\mathsf{E}[\ln p(x(t) \mid \Theta_z)]\right\} \sum_{z'} \alpha(z', t-1) \exp\left\{\mathsf{E}[\ln p(z \mid z', \{\Pi\})]\right\},$$
$$\beta(z,t) = \sum_{z'} \exp\left\{\mathsf{E}[\ln p(x(t+1) \mid \theta_{z'})]\right\} \beta(z', t+1) \exp\left\{\mathsf{E}[\ln p(z' \mid z, \{\Pi\})]\right\},$$

with the initial conditions $\alpha(z,1) = \exp\{\mathsf{E}[\ln p(z,1) + \ln p(x(1) \mid \Theta_z)]\}$ and $\beta(z,T) = 1\,\forall\, z$. These recursions are straightforwardly obtained by successive marginalization of Eq. (5.4.21) over $z(1), z(2), ..., z(t-1)$ (yielding $\alpha(z,t)$) and $z(T), z(T-1), ..., z(t+1)$ (yielding $\beta(z,t)$). The explicit derivations are provided in Appendix D.1 for completeness. These derivations also show that, in complete analogy to Section 2.2.1.1, the forward and backward messages yield the variational marginals via $q(z,t) \propto \alpha(z,t)\beta(z,t)$. The occurring expectations can be evaluated in closed form because of conjugacy between the variational distributions $q(\pi_z)$ and $q(\theta_z)$ and the corresponding likelihoods, see below.

Note that the observation likelihood is slightly different between the models for real-valued and angular data: as we utilize K priors $p(\theta_z)$, the periodic projection Eq. (5.3.14) is done *for each state* z:

$$x^z(t) \leftarrow x(t) - 2\pi \cdot \operatorname{sgn}(x(t) - \mu_{0,z}). \tag{5.4.25}$$

Accordingly, for the approximate von Mises model we set

$$p(x(t) \mid \theta_z) = p(x^z(t) \mid \theta_z). \tag{5.4.26}$$

TRANSITION DISTRIBUTIONS As discussed above, truncation of the variational posteriors to K states induces a partition on the index set. The exact prior can hence be written as a finite Dirichlet distribution, Eq. (5.4.23), allowing to readily compute

$$q\left(\pi_z \mid \{\eta_{z,j}\}_j\right) = \mathrm{Dir}(\pi_z \mid \eta_{z,1}, ..., \eta_{z,K}, \eta_{z,K+1}) \tag{5.4.27}$$

with the posterior concentration parameters

$$\eta_{z,j} = (\kappa\beta_{z'} + \delta_{z,z'}\xi) + \sum_t q(z', t, z, t-1) \text{ for } z' = 1, ..., K,$$
$$\eta_{z,K+1} = \kappa \cdot (1 - \sum_{z'=1}^{K} \beta_{z'j}).$$

OBSERVATION PARAMETERS As detailed in Section 5.3, we choose a NIW prior on the observation parameters $\theta_z = (\mu_z, \Sigma_z)$ to obtain conjugacy. This results in

$$q(\mu_z, \Sigma_z) = \mathrm{NIW}(\mu_z, \Sigma_z \mid \mu_{0,z}, \lambda_z, \Psi_z, \nu_z), \tag{5.4.28}$$

with

$$\begin{aligned} \mu_{0,z} &= \frac{\lambda\mu_0 + T\bar{x}_z}{\lambda + Q_z}, \quad \lambda_z = \lambda + Q_z, \quad \nu_z = \nu + Q_z, \\ \Psi_z &= \frac{\lambda(Q_z - T)}{\lambda + Q_z}\mu_0\mu_0^\top + M_z + S_z + \Psi_0, \end{aligned} \tag{5.4.29}$$

where $\mu_0, \lambda, \Psi, \nu$ are the parameters of the NIW-prior $p(\mu_z, \Sigma_z)$ and

$$\begin{aligned} \bar{x}_z &= \frac{1}{T}\sum_t x(t)q(z,t), \quad Q_z = \sum_t q(z,t), \\ M_z &= \frac{\lambda T}{\lambda + Q_z}(\bar{x}_z - \mu_{0,z})(\bar{x}_z - \mu_{0,z})^\top, \\ S_z &= \sum_t \left[q(z,t)x(t)x^\top(t) - \frac{\lambda + T}{\lambda + Q_z}\bar{x}_z\bar{x}_z^\top \right]. \end{aligned} \tag{5.4.30}$$

Note that the same remark as made regarding the latent state distributions applies here; in the approximate von Mises case, the data are wrapped around *each* state mean $\mu_{0,z}$ and accordingly enter Eq. (5.4.29) differently for each z, cf. Eq. (5.4.26).

TOP-LEVEL STICK-BREAKING MEASURE Setting up the transition distributions $p(\pi_z \mid \beta)$ via the defining DP-property Eq. (2.2.43) as $K+1$ Dirichlet distributions results in non-conjugacy between $p(\pi_z \mid \beta)$ and the stick-breaking measure $B \sim \mathrm{GEM}(\gamma)$ [182]. Thus, a closed-form update for β does not exist. It is customary to utilize a point estimate instead, $q(\beta) = \delta_{\beta^*}(\beta)$, rendering the expectation in Eq. (5.4.21) tractable [182]. While the optimum still has no closed-form solution, this allows us to utilize a gradient optimization scheme and update $\beta^* \leftarrow \beta^* + \omega\partial_{\beta^*}\mathsf{L}$. The gradient is found as

$$\begin{aligned} \partial_{\beta_z}\mathsf{L} &= \partial_{\beta_z}\left\{ \ln p(\beta) + \sum_{k=0}^{K} \mathsf{E}[\ln p(\Pi_k \mid \beta)] \right\} \\ &= 2\sum_{i \geq z} \ln \frac{1}{1 - \sum_{j<i}\beta_j} - (\gamma - 1)\sum_{i \geq k} \ln \frac{1}{1 - \sum_{j \leq i}\beta_j} \\ &\quad + \kappa\left(\sum_{k=1}^{K} \left[-\psi\left(\kappa\beta_z + \xi\delta_{k,z}\right) + \psi(\eta_{k,z}) - \psi(\eta_{k,K+1}) + \psi(\kappa\beta_{K+1}) \right] \right). \end{aligned} \tag{5.4.31}$$

The derivation of the gradients is provided in Appendix D.2. To set the step size ω, a back-tracking line-search algorithm is used as in Chapter 4: the step size is chosen as $\omega_i = 0.5^i$ and the current parameter β_z is updated as

$$\beta_z^{\text{new}} = \beta_z + \omega_i \partial_{\beta_z}\mathsf{L}\,.$$

If $\mathsf{L}[\beta^{\text{new}}] \geq \mathsf{L}[\beta]$, the update is accepted. Otherwise, we iterate and re-compute using the new step size ω_{i+1}.

5.5 RESULTS

The framework laid out above is applied to a range of different data sets: first, it is used on 2D HMM data to assess its functionality. It is then employed on continuous-valued SDE data generated from the standard three-well benchmark potential utilized in Chapter 4 to demonstrate its ability to identify a discrete, readily interpretable structure from all-continuous dynamics.

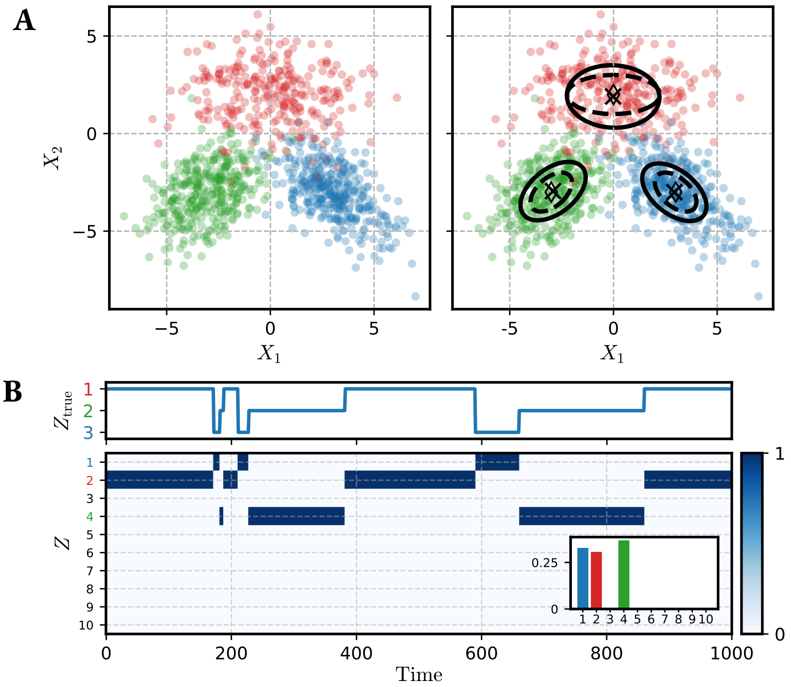


FIGURE 5.3: Inference on synthetic data generated from an HMM with three states and 2D Gaussian observation noise. For visualization purposes, only one of 10 simulated sequences is shown. **A**: Left: ground-truth data. Coloring indicates the corresponding latent state $z(t)$ for each data point $x(t)$. Right: inferred state assignments. Coloring is based on the MAP estimate of the latent state. Diamonds and dashed ellipses indicate the ground-truth means and covariances; the latter are represented by the 1-σ covariance ellipses. Solid crosses and ellipses indicate the inferred variational means $\mu_{0,z}$ and expected variational covariances $\Psi_z/(\nu_z - 3)$. **B**: Top: ground-truth latent sequence $z_{[1,T]}$. Bottom: reconstructed marginal probabilities. Every row $z \in \{1, 2, ..., 10\}$ indicates the posterior probability to be in state z at time point t, $q(z,t) \in [0,1]$. Inset: average state occupation $\langle q(z) \rangle = (IT)^{-1} \sum_{i,t} q(z^i, t)$ for all z.

Next, the applicability and limits of the vM approximation are studied on a 1D toy example. We then repeat the above procedure: we first show that the approximation works well on synthetic 2D von Mises data and then employ it on a standard MD benchmark dataset from the protein alanine dipeptide [132, 135, 187, 188]. Finally, the framework is applied to a large dataset from voltage clamp experiments on the viral potassium channel Kcv$_{\text{PBCV-1}}$ [189]. Experimental details and parameter settings are provided in Appendix D.3.

5.5.1 *Synthetic HMM data*

We specify a cyclic three-state HMM with transition probabilities

$$\Pi = \begin{pmatrix} 0.99 & 0.01 & 0 \\ 0 & 0.99 & 0.01 \\ 0.01 & 0 & 0.99 \end{pmatrix} \tag{5.5.32}$$

and uniform initial distribution $p(z,1)$. Each observation $x \in \mathbb{R}^2$ is drawn from a normal distribution with anisotropic covariances, $X \sim \mathcal{N}(\mu_z, \Sigma_z)$.

Ground truth	Inferred
$\mu_1 = \begin{pmatrix} 3 \\ -3 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$	$\mu_{0,1} = \begin{pmatrix} 2.94 \\ -3.01 \end{pmatrix}, \mathsf{E}[\Sigma_1] = \begin{pmatrix} 2.47 & 1.12 \\ 1.12 & 2.24 \end{pmatrix}$
$\mu_2 = \begin{pmatrix} -3 \\ -3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$	$\mu_{0,2} = \begin{pmatrix} -2.91 \\ -2.95 \end{pmatrix}, \mathsf{E}[\Sigma_2] = \begin{pmatrix} 2.40 & -1.11 \\ -1.11 & 2.18 \end{pmatrix}$
$\mu_3 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$	$\mu_{0,3} = \begin{pmatrix} -0.02 \\ -1.49 \end{pmatrix}, \mathsf{E}[\Sigma_3] = \begin{pmatrix} 4.98 & -0.14 \\ -0.14 & 2.56 \end{pmatrix}$

TABLE 5.1: Parameters learned from 2D, three-state HMM data with Gaussian emissions.

As shown in Fig. 5.3 B, the method accurately recovers three latent states. The state sequence is also identified correctly, as can be seen from the inferred marginals $q(z, t)$ of one latent state sequence. In particular, the maximum a posteriori (MAP) assignment $z(t)$ of each data point $x(t)$ defined via

$$z^{\mathrm{MAP}}(t) = \arg\max_z q(z, t) \tag{5.5.33}$$

precisely matches the corresponding ground truth.

Accordingly, also the inferred posterior means $\mu_{0,z}$ and expected covariances (black crosses and solid ellipses in Fig. 5.3)

$$\mathsf{E}\left[\Sigma_z\right] = \frac{\Psi_z}{\nu_z - n - 1}, \tag{5.5.34}$$

with $n = 2$ the dimensionality of the system, faithfully resemble their ground-truth counterparts (diamonds and dashed ellipses in Fig. 5.3; the expected values of the inferred observation parameters are summarized in Table 5.1 below). Notice that the inferred state *labels* of course do not necessarily correspond to the ground-truth labels: this is an interpretation to be done by the experimenter after convergence of the model. For illustrative purposes, one trajectory over all K states is shown in Fig. 5.3.

Similarly, the transition distributions accurately recover the prior transition matrix as can be seen by comparing the expected transition probabilities

$$\mathsf{E}[\Pi_{zz'}] = \frac{\eta_{z,z'}}{\sum_{z'} \eta_{z,z'}}$$

between the three mainly-occupied states (cf. Fig. 5.3) to Eq. (5.5.32), see Fig. 5.4.

5.5.2 *Stochastic dynamics in a 2D potential*

Next, the method is applied to data generated from the 2D benchmark potential given in Section 4.5.2, which is a standard problem in MSM analysis [131, 136, 190]. The dynamics are given as before via an SDE evolving in the potential landscape Eq. (4.5.26).

The potential, together with the inferred state means $\mu_{0,z}$ and expected covariances $\mathsf{E}[\Sigma_z]$, is shown in Fig. 5.5. This reconstruction captures the essential system features well. The locations

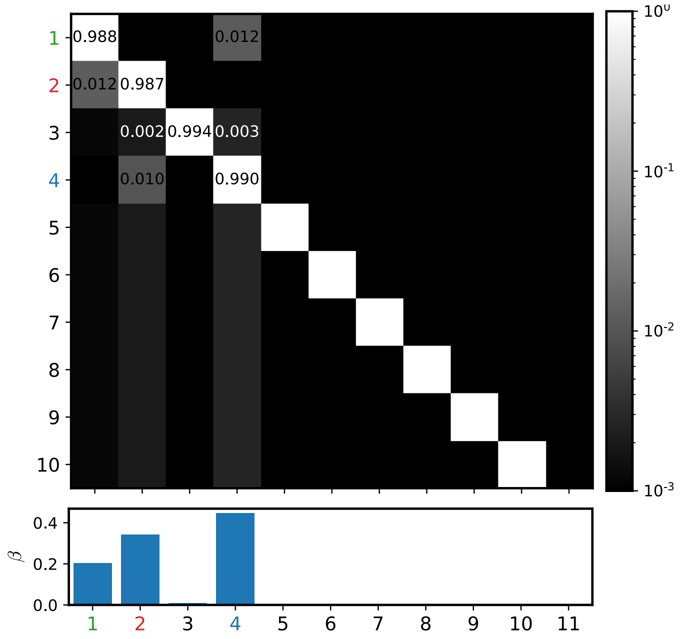


FIGURE 5.4: Top: Expected values of the variational transition probabilities $\mathsf{E}[\Pi_z]$. Bottom: point-estimate of the top-level stick-breaking measure β. Note how the structure of β is reflected in $\mathsf{E}[\Pi_z]$.

Ground truth	$\mu_1 = \begin{pmatrix} 0 \\ 1.5 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
Inferred	$\mu_{0,1} = \begin{pmatrix} -0.07 \\ 1.09 \end{pmatrix}, \quad \mu_{0,2} = \begin{pmatrix} -0.98 \\ -0.03 \end{pmatrix}, \quad \mu_{0,3} = \begin{pmatrix} 0.96 \\ -0.02 \end{pmatrix}$

TABLE 5.2: Parameters learned from three-well diffusion data

of the potential minima are accurately recovered; notably, the two deeper minima at $x_2 = 0$ are identified with higher precision than the shallow one at the top. This is not surprising, as due to its shallowness, the systems spends a comparably short time in this region, generating only little respective evidence.

The inferred sequence of metastable states yields accordingly plausible results, which can be seen by comparing the components of the true continuous process to the inferred discrete switching process.

5.5.3 *Testing the approximate von Mises model: 1D toy example*

To provide an intuition about the approximate von Mises model and to demonstrate the effect of data concentration on its validity, we generate synthetic data from a 1D vM Eq. (5.3.12) and infer the respective approximate posterior *without* an underlying switching process. The results are shown in Fig. 5.6: It is immediately clear that for sufficiently high concentration values, the approximation error (indicated by the red shaded area outside the unit circle in Fig. 5.6, left) becomes negligible. The approximate vM hence allows for staightforward debugging: as long as the probability assigned to the area outside the unit circle is small, the approxmation can be

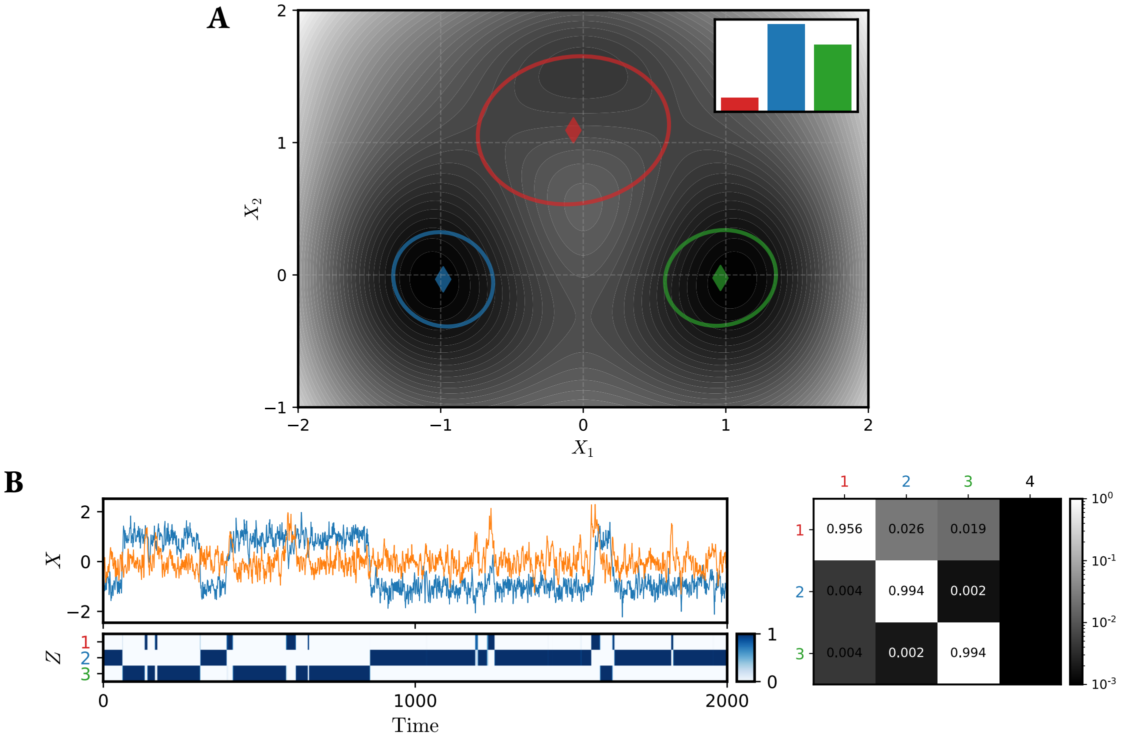


FIGURE 5.5: Inference of metastable states of 2D SDE dynamics. **A**: The heatmap shows the potential landscape used to simulate the continuous dynamics (brighter colors indicate higher values). Colored diamonds and ellipses indicate the inferred variational means and 1-σ ellipses of the expected variational covariances $\Psi_z/(\nu_z - 3)$, cf. Eq. (5.5.34). Inset: average state occupation $\langle q(z)\rangle = (IT)^{-1}\sum_{i,t} q(Z^i(t) = z)$ for all 3 identified metastable states. **B**: Left: Part of one simulated trajectory (top; X_1-component in blue, X_2-component in orange) and the corresponding latent sequence reconstruction (bottom). Right: Expected values of the variational transition probabilities $\mathsf{E}[\Pi_k]$. Label colors indicate the variational modes.

assumed to be valid. Vice versa, it deteriorates if this probability becomes non-negligible. We accept this error in the observation model to arrive at a tractable expression.

5.5.4 *Synthetic HMM data with angular observations*

We now apply the method to an angular setting analogous to that of Section 5.5.1: we generate data from the same HMM, but employ a von Mises observation model. Two of the three latent states generate independent 1D von Mises observations along each dimension, cf. Eq. (5.3.12). The third state includes correlations and generates observations from the bivariate von Mises distribution Eq. (5.3.10) following the sampling scheme of [180], which is briefly reiterated for convenience: a sample (ϕ, ψ) is obtained by subsequently sampling ψ from the marginal $p(\psi)$ and ϕ from the conditional $p(\phi \mid \psi)$. By marginalizing Eq. (5.3.10) over ϕ, one obtains

$$p(\psi) = c(\lambda_1, \lambda_2, \lambda_3) 2\pi I_0\left(\sqrt{\lambda_1^2 + \lambda_3^2 - 2\lambda_1\lambda_2\cos(\psi - \nu)}\right)\exp(\lambda_2\cos(\psi - \nu)),$$

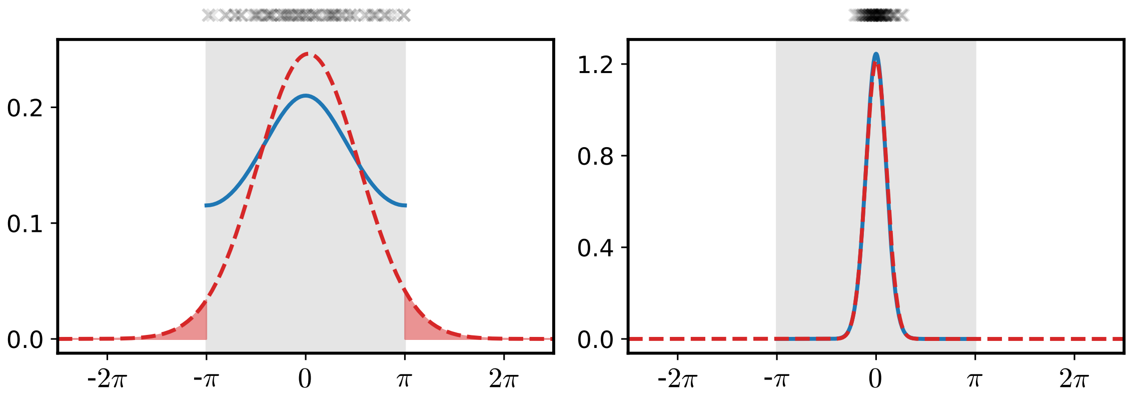


FIGURE 5.6: Comparison of a ground-truth one-dimensional von Mises with the posterior obtained via the approximate von Mises model. Left: inferred approximate von Mises (red dashed line) for $N = 1000$ data points generated from a low-concentration von Mises (cf. Eq. (5.3.12), $p(\cdot \mid \nu, \lambda)$ with $\nu = 0, \lambda = 0.3$, blue line). 10% of data points are shown above the plot. Grey shaded area: unit circle. Red shaded area: probability mass outside the unit circle boundaries. Right: inferred approximate von Mises for $N = 1000$ data points generated from a high-concentration von Mises, $\lambda = 10$. No significant probability mass is placed outside the unit circle.

from which a sample is drawn via rejection sampling with a simple one-dimensional vM as a proposal distribution. The conditional density $p(\phi \mid \psi)$ is similarly found to be vM [191]:

$$p(\phi \mid \psi) = p_{\text{vM}}\left(\phi \,\middle|\, \arctan\left(\frac{-\lambda_3 \sin(\psi - \nu)}{\lambda_1 - \lambda_3 \cos(\psi - \nu)}\right), \sqrt{\lambda_1^2 + \lambda_3^2 - 2\lambda_1\lambda_3 \cos\psi}\right).$$

Similar to the non-angular case, we generate a certain degree of overlap between the individual distributions and include observations wrapping around the period boundary $2\pi \to 0$.

Ground truth	Inferred
$(\mu_1^1, \lambda_1^1) = (0, 20), (\mu_1^2, \lambda_1^2) = (-0.9\pi, 3)$	$\mu_{0,1} = (0, -0.9\pi)^\top$
$(\mu_2^1, \lambda_2^1) = (-0.4\pi, 3), (\mu_2^2, \lambda_2^2) = (-0.6\pi, 30)$	$\mu_{0,2} = (-0.4\pi, -0.6\pi)^\top$
$(\mu_3^1, \mu_3^2, \lambda_3^1, \lambda_3^2, \lambda_3^3) = (0.98\pi, 0.5\pi, 10, 5, 2)$	$\mu_{0,3} = (0.97\pi, 0.51\pi)^\top$

TABLE 5.3: Parameters learned from 2D, three-state von Mises data with the approximate vM model. Left: ground-truth tuples of positions and concentration parameters. Note that the first two states are characterized by one 1D vM per dimension (indicated by superscripts) and the third state by one 2D vM according to Eq. (5.3.10).

As shown in Fig. 5.7, the von Mises approximation accurately recovers the ground-truth means, see also Section 5.5.4. Since the ground-truth data are generated by *true* von Mises distributions, no ground-truth covariance matrices exist; hence, the inferred covariances cannot be directly compared to them. However, we can assess by comparison with the plotted data that the von Mises approximation produces accurate estimates. Our projection method Eq. (5.3.14) also enables sensible and accurate periodic continuations across the period boundaries.

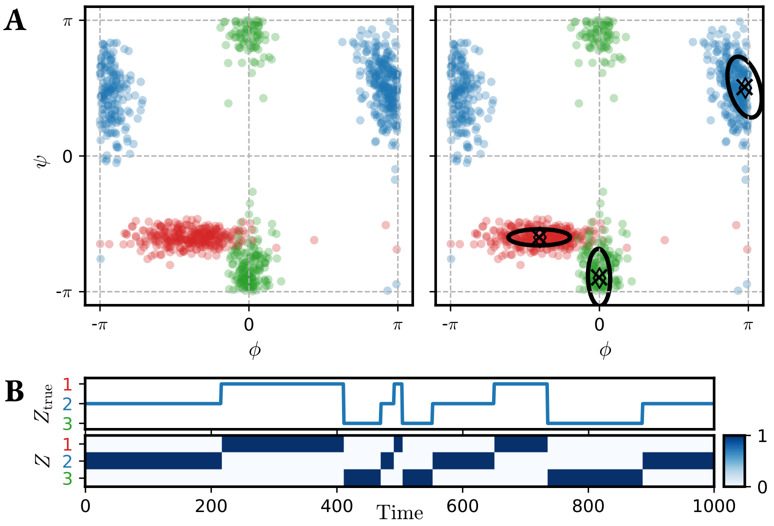


FIGURE 5.7: Inference on synthetic von Mises data utilizing the approximate von Mises model. For visualization purposes, only one of 10 simulated sequences is shown. **A**: Left: ground-truth data. Coloring indicates the corresponding latent state $z(t)$ for each data point $x(t)$. Right: inferred state assignments. Coloring according to the MAP estimate of the latent state. Diamonds indicate the ground-truth means. Black crosses and ellipses indicate the inferred variational means $\mu_{0,z}$ and expected variational covariances $\Psi_z/(\nu_z-3)$. **B**: Top: ground-truth latent sequence $z_{[1,T]}$. Bottom: reconstructed marginal probabilities. Every row $z \in \{1, 2, ..., 10\}$ indicates the posterior probability to be in state z at time point t, $q(z,t)$.

5.5.5 *MD simulation data: alanine dipeptide*

After benchmarking the approximate von Mises model, we now apply it to real-world MD simulation data. We utilize a data set of the molecule alanine dipeptide provided with the `pyemma` package [159] consisting of $I = 3$ independent trajectories of length $T = 250000$ ps with a time step of 1 ps. Alanine dipeptide is a commonly-used model system in computational biology [192–194]. This 2D dataset describes the molecule dynamics in terms of two backbone torsion angles (ϕ, ψ). Inspection of the raw data, Fig. 5.8, shows that the simulation exhibits metastable dynamics, highlighting the relevance of the metastability assumption at the outset.

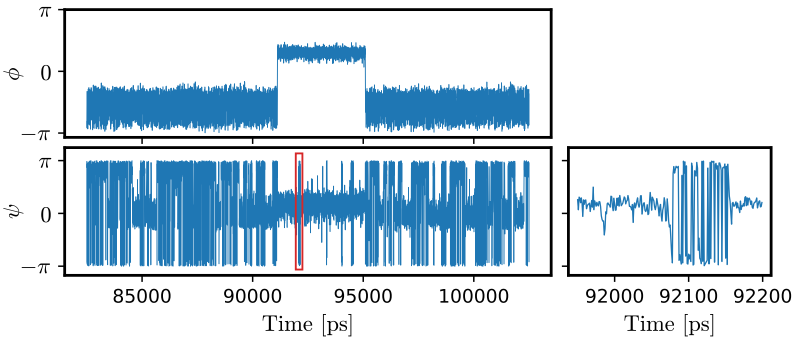


FIGURE 5.8: Excerpts of alanine dipeptide MD simulations. Left: excerpt of 20000 ps of one trajectory. Roughly between 90000 and 95000 ps a switching event occurs that can be identified as a transition to the states α_L and α_D (cf. Fig. 5.9). Right: close-up of the red box shown in the left plot. Note that what looks like high-frequency transitions is simply an artifact of the wrap-around across the period boundary.

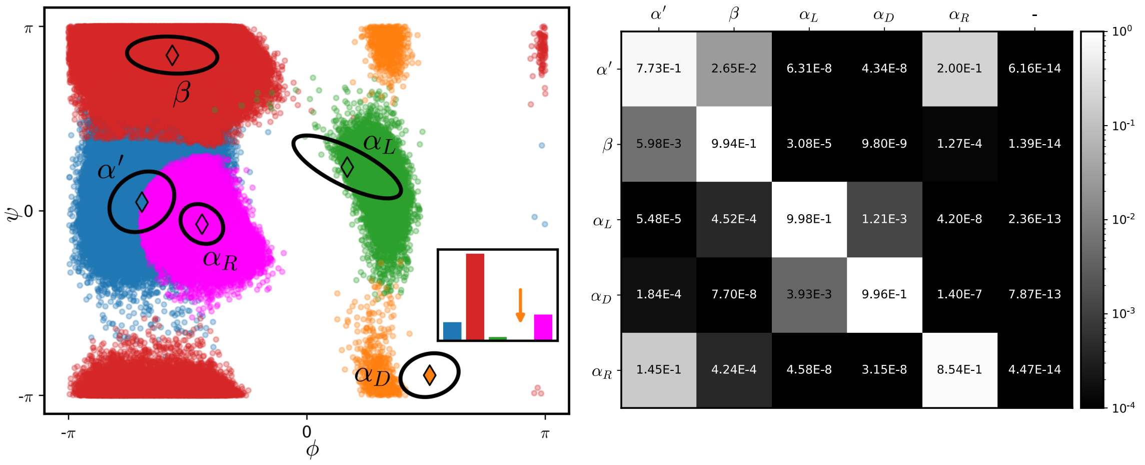


FIGURE 5.9: Metastable dynamics of alanine dipeptide. Left: conformational states. Coloring indicates the MAP state assignment. The inset shows the relative proportions of all 5 states occurring in the data. State 4 (orange arrow) only occurs $< 1\%$. As before, colored diamonds and ellipses represent the variational means and covariances. Annotations refer to known α-helix and β-sheet conformations, see for instance [192, 196]. Right: expected transition probabilities. Each row shows the transitions probabilities $\mathsf{E}[\Pi_k]$ from one of the five found states to all others, including the "rest" of the state space (cf. Eq. (2.1.2)) indicated by "$-$". Note that the transitions $\alpha' \to \alpha_R$ and $\alpha_R \to \alpha'$ are approximately equal.

The conformational landscape of alanine dipeptide exhibits a complex fine structure. Due to its widespread adoption in the field, a host of computational frameworks has been applied to this dataset: typically, partitionings between three and six different states are obtained [132, 135, 187, 188, 195]. As shown in Fig. 5.9, our framework identifies five different states consistent with this literature. By comparison to in-depth MD studies of alanine dipeptide [192, 196], these five states can be matched to known α-helix and β-sheet conformations.

As an aside, note that the transition probabilities between the two states α' and α_R are found to be almost symmetrical. It would hence be a valid interpretation of the model results that these two states could be lumped together. This would similarly be in line with the literature [192, 196].

5.5.6 *Electrophysiological single-molecule ion channel data*

Finally, we apply our method to voltage-clamp data of the viral potassium channel $\mathrm{Kcv}_{\mathrm{PBCV-1}}$. The wild-type channel switches between an "open" and a "closed" state; mutation of the last amino acid to histidine, however, leads to the appearance of sublevels between "open" and "closed" [197]. Here, we utilize our method to quantify these sublevels. The data are obtained using the planar lipid bilayer technique as detailed in [142]: the applied voltage is 160 mV at pH=6 and data are recorded at 5 kHz over a time span of $T = 60$ s. Half of the data are discarded due to apparent drift, see Fig. 5.10.

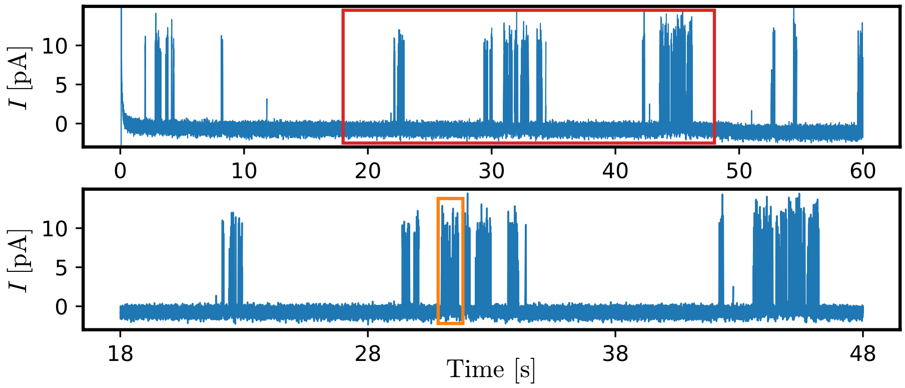


FIGURE 5.10: Raw data of the electrophysiological experiment. Top: Full trajectory. Bottom: close-up of the red box in the top plot; this is the data the algorithm is run on. The orange box is shown with the inferred states in Fig. 5.11.

Despite the high noise level, the inferred latent sequence in Fig. 5.11 shows a highly plausible switching behavior. Three distinct states can be identified: the "closed" and "open" states known from the wild-type as well as one intermediate, subconductive state; see also Table 5.4. We hence establish that the histidine mutation yields *one* novel channel conformation not attained by the wild-type.

Inferred means	$\mu_c = -0.698\,\text{pA}, \quad \mu_i = 2.33\,\text{pA}, \quad \mu_o = 7.63\,\text{pA}$
Inferred covariances	$\mathsf{E}[\Sigma_c] = 0.17\,\text{pA}^2, \quad \mathsf{E}[\Sigma_i] = 4.28\,\text{pA}^2, \quad \mathsf{E}[\Sigma_o] = 1.15\,\text{pA}^2$

TABLE 5.4: Parameters learned from switching ion channel data from viral potassium channel $\text{Kcv}_{\text{PBCV}-1}$.

Note that one full optimization run only took $\sim 25\,\text{s}$ for a sequence of 1.5×10^5 time points, which is orders of magnitude faster than the sampling algorithm proposed in [152] for analysis of such trajectories. Notably, conventional methods of trajectory segmentation [198] require both the pre-specification of the number of conformational states as well as their conductivity values, which the presented method does not.

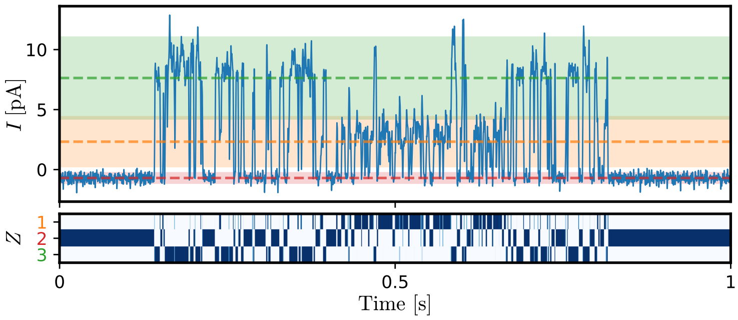


FIGURE 5.11: Conformational states of the viral potassium channel $\text{Kcv}_{\text{PBCV}-1}$. For illustrative purposes, we constrain the figure to an interval of $T = 1\,\text{s}$ rather than showing the full trajectory, cf. Fig. 5.10. Top: Measured current I over time. Dashed lines represent the inferred posterior means, shaded regions the expected variational standard deviation. Green: "open" state. Red: "closed" state. Orange: intermediate state. Bottom: inferred latent sequence.

5.6 SUMMARY

The nonparametric framework presented in this chapter offers a generative modeling approach for Bayesian inference of metastable conformational states from experimental and simulation data. This allows the user to leave the number of structural states unspecified a priori and learn it from the data, which is beneficial, as this number is not known in advance in typical experimental and computational settings. The HDP-HMM is a generalization of the widely applied MSM framework: in contrast to the MSM method, one neither (i) needs to pre-process the data via discretization and temporal thinning to re-establish Markovianity, nor (ii) manually select the number of metastable states. Importantly, the proposed HDP-HMM approach does not deteriorate the temporal resolution of the data.

As demonstrated, this model is able to reliably identify metastable states - their number has been sensibly established in all experiments. The application to the stochastic particle dynamics data highlights the utility of this model on purely continuous data as generated, e.g., by MD. It hence achieves the central goal of modeling the continuous dynamics via readily interpretable discrete conformational states.

Furthermore, in an attempt to adapt the HDP-HMM to a setting often encountered in MD, a computationally tractable approximation to the von Mises distribution was proposed, as MD data are frequently specified via dihedral angles; this yielded accurate results. Results obtained on a canonical benchmark data set of alanine dipeptide are consistent with the existing literature. We emphasize that this benchmark problem, while consisting of relatively short trajectories compared to MD standards, requires the use of variational inference methods, as MCMC-type sampling schemes are computationally intractable. This point was also highlighted via inference on experimental voltage clamp data, where existing methods all resort to sampling schemes [152–155] and hence require runtimes on much longer time scales than the presented framework. For a discussion of possible extensions or other applications of the presented framework, the reader is referred to Chapter 6.

6 CONCLUSIONS

6.1 SUMMARY OF CONTRIBUTIONS

In Chapter 3, an MCMC framework for MJP-SSDE processes was proposed. The framework generates samples from the exact conditional posterior measures and yields accurate results in particular for the latent system dynamics. Parameter estimation works well for 1D systems; in 2D, more structured prior distributions would aid performance. Algorithm runtime may be challenging, however, as a system of several ODEs and stochastic integrals has to be solved for each sample.

To address this computational issue, a variational approximation for the same system class was put forward in Chapter 4. It was shown that under conditions of metastability, the approximation produces accurate estimates of the system dynamics as well as the parameters. If the metastability criterion is not fulfilled, however, the VI approach fails to recover the true latent dynamics.

Chapter 5 presented a Bayesian nonparametric VI approach to the problem of conformational molecule switching. The method faithfully identifies the number of conformations in the examined problems and addresses two shortcomings of the classic MSM framework as well as the need for a computationally more efficient algorithm than HDP-HMM sampling approaches.

6.2 OUTLOOK

The results presented in this thesis open several potential avenues for further research. Three main areas will be discussed in the following:

1. methodological improvement,
2. biological application, and
3. conceptual development.

1. METHODOLOGICAL IMPROVEMENT An effort to improve on the presented MCMC method could build on particle filtering and smoothing [109, 199]: the Gibbs sampler could be used within a particle smoother framework, where the conditional posterior measures serve as proposal distributions. As these are exact measures, the generated samples - albeit being approximate due to discretization - should yield high likelihood weights. This could aid in ameliorating the slow mixing properties of the Gibbs sampler [200]. The same goal could, on the other hand, also be pursued by extending the presented approach to a Hamiltonian Monte Carlo scheme [201]. Vice versa, if the framework is altered by employing another SSDE instead of the simple linear one utilized in this work, a particle smoothing approach could be used to generate respective samples of the diffusion.

Note that apart from the discussed linear-Gaussian case, closed-form solutions to the FPE also exist, e.g., for SDEs driven more generally by the gradient of a potential fulfil certain requirements to allow for a closed-form solution [202]. It would be interesting to explore the potential and limitations of such an approach in applying the method to purely continuous dynamics, akin to the one presented in Section 5.5.2. This could be complemented by and compared to an approach utilizing neural networks to solve the FPE [203–206], which would allow for even more flexibility.

2. BIOLOGICAL APPLICATIONS An interesting application of the MJP-SSDE model would be the “flickering” of ion channels [207]. This phenomenon occurs when an ion channel enters a regime of rapid switching between states which is too fast to be resolved by typical measurement devices. It results in measurements of effective currents of intermediate strength between the levels of the two rapid-switching states. The idea is that the measurement of a rapid and large change in current should be enough to increase the likelihood of a switch; it should not be necessary to require the process to relax to its local steady state. This however first requires methodological improvements (such as the ones described above) to ensure fast mixing and increased numerical stability, due to the large changes in small time intervals on the order of magnitude of the SDE discretization time step.

The MJP-SSDE model furthermore lends itself to optimal control problems, which have for switching systems been intensely studied in recent years [208–211]. For instance, it could first be attempted to gauge the parameters of a genetic switching system as the one presented in Chapter 3 accurately enough so as to be able to devise control schemes to keep its TX-TL activity and the resulting fluorescence level in a predefined region. This could directly be compared to wet-lab experiments, cf. Section 3.4. This setup could be further extended to allow for inter-molecule communication, where the control signal of one cell is generated by another cell in its vicinity, which itself is subject to some external signal.

3. CONCEPTUAL DEVELOPMENT Finally, the work presented points towards a unified framework for the modeling of stochastic hybrid systems: in principle, one would like to have a nonparametric, recurrent hybrid process model with complex observation likelihood functions and non-exponential sojourn times.

First, the nonparametric formulation of the problem from Chapter 5 needs to be transferred to the continuous-time regime. Prior work that can be built upon already exists, see, e.g., [100, 212, 213].

The term "recurrent" here refers to a generalization of the proposed MJP-SSDE where the continuous process Y feeds back to the switching process Z. This makes sense in particular for the conformational switching of molecules, as the conformations are defined by the relative continuous atom positions of the molecules. Changes in these positions should hence influence the transition probability between global conformations. Analogous concepts have recently been developed in the discrete time regime [93, 94].

For molecular switching specifically, the assumed Gaussian (or von Mises) observation likelihoods are not expressive enough to capture complex conformational energy landscapes. To address this issue, the latent hybrid process could be combined with recent neural network approaches to continuous-time processes [214]. Akin to the classic variational auto-encoder, this could also achieve an efficient encoding to lower dimensions [215]. We note that in the field of Markov state modeling, approaches to this challenge have been proposed recently [132, 177]. None of these proposals however builds on nonparametric formulations. Another approach from machine learning combines conventional probabilistic models with complex likelihood functions in a modular way - compromising, however, the lower-bound property of the ELBO [97]. While Chapter 4 pursues a similar approach, this is potentially more problematic in the nonparametric case, as less knowledge about the system at hand is available from the outset. This makes it harder to justify an approximation to the ELBO that does not preserve a lower bound.

Finally, a sensible extension would be semi-Markovian models, in which the transitions between different states are still Markovian, but the sojourn times within each state may be non-exponentially distributed: due to the exponential sojourn time distribution, fast transitions are favored over slow ones, which may be too strong an assumption for general conformational switching. Similar analyses have already been carried out for ion channel data and might hence help to get a more detailed understanding of complex switching dynamics [216]. The challenge here is to obtain computationally tractable inference schemes; see, e.g., [217]. Notice that similar ideas are also already being exploited for lumping in conventional MSM settings [164] and alleviating the lag-time issue of MSMs [160].

APPENDICES

A

APPENDIX A: BACKGROUND

A.1 PROBABILITY THEORY AND NOTATION

In the following, some basic probability theory will be reviewed. For extensive treatments, see, e.g., [13, 101, 218].

PROBABILITY SPACES A probability space $(\Omega, \mathcal{F}, \mathsf{P})$ consists of a non-empty set Ω (the *sample space*), a *sigma algebra* $\mathcal{F} \subseteq 2^{\Omega}$ and the *probability measure* $\mathsf{P} : \mathcal{F} \to [0, 1]$.

RANDOM VARIABLES, EXPECTATIONS AND DISTRIBUTIONS A random variable Z is a function $Z : \Omega \to \mathcal{Z}$ from the sample space to some measurable space $\mathcal{Z}$. Throughout this thesis - unless stated otherwise - random variables are denoted by roman upper case letters such as Z and the corresponding realizations by lower case letters, $Z = z$.

The expectation of a random variable is defined as its integral with respect to some probability measure,

$$\mathsf{E}[Z] = \int_{\mathcal{Z}} Z \mathrm{d}\mathsf{P} = \int_{\mathcal{Z}} Z(z) \mathsf{P}(Z \in \mathrm{d}z),$$

where we introduced a common (differential) notation for measures as $\mathsf{P}(Z \in \mathrm{d}z)$, see [13]. For brevity, we will write

$$\mathsf{E}[Z] = \int_{\mathcal{Z}} z \, \mathsf{P}(Z \in \mathrm{d}z).$$

If $\mathcal{Z} \subseteq \mathbb{N}$ and the measure P is absolutely continuous with respect to the counting measure, we can write

$$\mathsf{E}[Z] = \sum_{z \in \mathcal{Z}} z p(z),$$

with the probability mass function (PMF)

$$p(z) = \mathsf{P}(Z = z).$$

If, on the other hand, $\mathcal{Z} \subseteq \mathbb{R}$ and the measure is absolutely continuous with respect to the Lebesgue measure, we have

$$\mathsf{E}[Z] = \int_{z \in \mathcal{Z}} z p(z) \mathrm{d}z,$$

with the probability density function (PDF)

$$p(z) = \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}z}.$$

For two random variables Y and Z, we furthermore have

$$\mathsf{E}[Y] = \mathsf{E}[\mathsf{E}[Y \mid Z]]$$

with the *conditional* expectation $\mathsf{E}[Y \mid Z]$, in which the expectation is only taken over Y; that is, a conditional expectation is still itself a random variable. In measure notation, we write

$$\mathsf{E}\left[\mathsf{E}[Y \mid Z]\right] = \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} y \, \mathsf{P}(Y \in \mathrm{d}y \mid Z = z) \right) \mathsf{P}(Z \in \mathrm{d}z).$$

For conciseness, we abbreviate the conditioning set throughout as

$$\mathsf{P}(Y \in \mathrm{d}y \mid Z = z) = \mathsf{P}(Y \in \mathrm{d}y \mid z).$$

Regarding distributions, we write "$\sim$" for "is distributed according to". We discern between standard measures and their respective PDFs by overloading: "the random variable X is distributed according to the Gaussian measure with mean μ and covariance Σ" is written as

$$X \sim \mathcal{N}(\mu, \Sigma),$$

and the respective density function

$$\begin{aligned} p(x) &= \mathcal{N}(x \mid \mu, \Sigma) \\ &= \det(2\pi\Sigma)^{-1/2} \exp\left\{ -\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}. \end{aligned}$$

Between two different measures P, Q on the same space $\mathcal{Z}$, the Kullback-Leibler (KL) divergence (also known as *relative entropy*) is defined as

$$\mathsf{D}_{\mathrm{KL}}[\mathsf{P} \mid\mid \mathsf{Q}] := \mathsf{E}\left[\ln \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\mathsf{Q}}\right] = \int_{\mathcal{Z}} p(z) \ln \frac{p(z)}{q(z)} \mathrm{d}z$$

where the last equality requires the measures to admit respective densities $p(z), q(z)$.

STOCHASTIC PROCESSES For a stochastic process $Z := \{Z(t) : t \in \mathcal{I}\}$, we denote with $p(z,t)$ the density function of z at time point t,

$$\mathsf{P}(Z(t) \in \mathcal{A}) = \int_{\mathcal{A}} p(z,t)\mathrm{d}z,$$

with some set $\mathcal{A} \subseteq \mathcal{F}$. This can equivalently be expressed via a differential expression,

$$p(z,t) := \partial_z \mathsf{P}(Z(t) \leq z).$$

This also holds for multiple time points, e.g.,

$$p(z,t,z',t') := \partial_z \partial_{z'} \mathsf{P}(Z(t) \leq z, Z(t') \leq z').r,$$

and for conditional densities,

$$p(z',t' \mid z,t) := \partial_{z'} \mathsf{P}(Z(t') \leq z' \mid Z(t) = z).$$

As an aside, note that in a mathematically rigorous treatment of stochastic processes, the conditional expectation is at center stage. Recommendable reads on this are [82, 219].

As an example of a stochastic process, consider standard Brownian motion, which is used frequently throughout the thesis. Brownian motion $W(t)$ is defined by four properties:

1. $W(t) = 0$
2. is has independent increments
3. $W(t) - W(s) \sim \mathcal{N}(0, t-s), \quad 0 \leq s \leq t$
4. it is continuous almost surely.

A.2 DERIVATION OF THE FOKKER-PLANCK EQUATION

Start with the Itô formula (2.1.22) for some stochastic process X with generator $\mathcal{L}_t$,

$$\mathrm{d}f(X(t),t) = [\partial_t + \mathcal{L}_t] f(X(t),t)\mathrm{d}t + \mathrm{d}\mathcal{M}(t),$$

and assume a time-independent function $f(x,t) = f(x)$. We write (in integral form)

$$f(X(T)) - f(X(0)) = \int_0^T \mathcal{L}_t f(X(t),t)\mathrm{d}t + \mathcal{M}(T) \tag{A.2.1}$$

and take the expectation, yielding

$$\begin{aligned}
\mathsf{E}[f(X(T))] - \mathsf{E}[f(X(0))] &= \mathsf{E}\left[\int_0^T \mathcal{L}_t f(X(t))\mathrm{d}t\right] \\
&= \int_0^T \mathsf{E}\left[\mathcal{L}_t f(X(t))\right] \mathrm{d}t \\
&= \int_0^T \int p(x,t)\mathcal{L}_t f(x,t)\mathrm{d}x\mathrm{d}t \\
&= \int_0^T \int f(x,t)\mathcal{L}_t^\dagger p(x,t)\mathrm{d}x\,\mathrm{d}t. \qquad \text{(A.2.2)}
\end{aligned}$$

The last equality makes use of the L_2-adjoint $\mathcal{L}_t^\dagger$ of the operator $\mathcal{L}_t$: with

$$\langle p, \varphi \rangle := \int p(x,t)\varphi(x,t)\,\mathrm{d}x$$

and an appropriate test function φ,

$$\langle p, \mathcal{L}\varphi \rangle = \langle \mathcal{L}^\dagger p, \varphi \rangle.$$

The left-hand side can be interpreted as functions of t; thus,

$$\begin{aligned} \mathsf{E}[f(X(T))] - \mathsf{E}[f(X(0))] &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\,\mathsf{E}[f(X(t))]\mathrm{d}t \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\int f(x)p(x,t)\mathrm{d}x\mathrm{d}t \\ &= \int_0^T \int f(x)\partial_t p(x,t)\mathrm{d}x\mathrm{d}t. \end{aligned} \tag{A.2.3}$$

As f is arbitrary, the KFE follows,

$$\partial_t p(x,t) = \mathcal{L}^\dagger p(x,t). \tag{A.2.4}$$

A.3 DERIVATION OF THE KOLMOGOROV BACKWARD EQUATION

Let now specifically

$$f(x,t) := \mathsf{E}[\varphi(X(T)) \mid X(t) = x] \tag{A.3.5}$$

with some arbitrary function φ. In this case, we find

$$\begin{aligned} &\mathsf{E}[f(x,T) \mid X(0) = x] = \mathsf{E}[\varphi(X(T)) \mid X(0) = x] = \varphi(x) \\ &\mathsf{E}[f(x,0) \mid X(0) = x] = \varphi(x). \end{aligned}$$

Accordingly,

$$\mathsf{E}[f(X(T),T) \mid X(0) = x] - \mathsf{E}[f(X(0),0) \mid X(0) = x] = 0$$

and Itôs lemma then reads

$$0 = \int_0^T \mathsf{E}[(\partial_t + \mathcal{L}_t)\, f(X(t),t) \mid X(0) = x]\mathrm{d}t$$

Taking the limit on the integral and applying the mean value theorem shows that indeed the integrand itself has to vanish:

$$0 = \lim_{T\to 0}\frac{1}{T}\int_0^T \mathsf{E}[(\partial_t + \mathcal{L}_t)\, f(X(t),t) \mid X(0) = x]\mathrm{d}t \tag{A.3.6}$$

$$\Rightarrow \partial_t f(x,t) = -\mathcal{L}_t f(x,t). \tag{A.3.7}$$

This is the Kolmogorov backward equation (KBE); the KBE for the transition density in particular follows upon insertion of Eq. (A.3.5).

B

APPENDIX B: MARKOV CHAIN MONTE CARLO FOR HYBRID SYSTEMS

B.1 THE HYBRID MASTER EQUATION

B.1.1 *Derivation of the prior hybrid master equation*

To derive the HME, we assume for simplicity that $Z(t) \in \mathcal{Z} \subseteq \mathbb{N}$ and $Y(t) \in \mathcal{Y} \subseteq \mathbb{R}$. The multivariate case $\mathcal{Y} \subseteq \mathbb{R}^n$ can be derived analogously.

Following [102], we utilize the Chapman-Kolmogorov equation (2.1.11):

$$\begin{aligned}
p(y, z, t+h \mid \mathcal{X}) &= \sum_{z'} \int_{\mathcal{Y}} p(y, z, t+h \mid y', z', t, \mathcal{X}) p(y', z', t \mid \mathcal{X}) \, \mathrm{d}y' \\
&= \sum_{z'} \int_{\mathcal{Y}} p(y, t+h \mid z, t+h, y', z', t, \mathcal{X}) p(z, t+h \mid y', z', t, \mathcal{X}) \\
&\qquad \cdot p(y', z', t \mid \mathcal{X}) \, \mathrm{d}y',
\end{aligned}$$

where $\mathcal{X} = \{y(0), z(0)\}$.

We utilize the expansion of the transition density

$$p(z, t+h \mid y', z', t, \mathcal{X}) = \delta_{z'z} + \Lambda^{y'}_{z'z}(t) h + o(h)$$

with $\lim_{h \to 0} \frac{1}{h} p(z, t+h \mid y', z', t, \mathcal{X}) =: \Lambda^{y'}_{z'z}(t)$. As we will let $h \to 0$ at the end, we omit terms of $o\,(h)$ in the following. Inserting this expansion into the above expression yields

$$\begin{aligned}
&p(y, z, t+h \mid \mathcal{X}) \\
&= \sum_{z'} \int_{\mathcal{Y}} p(y, t+h \mid z, t+h, y', z', t, \mathcal{X}) \left(\delta_{z'z} + \Lambda^{y'}_{z'z}(t) h \right) p(y', z', t \mid \mathcal{X}) \, \mathrm{d}y' \\
&= \int_{\mathcal{Y}} p(y, t+h \mid z, t+h, y', z, t, \mathcal{X}) p(y', z, t \mid \mathcal{X}) \, \mathrm{d}y' \\
&\quad + \sum_{z'} \int_{\mathcal{Y}} p(y, t+h \mid z, t+h, y', z', t, \mathcal{X}) \Lambda^{y'}_{z'z}(t) p(y', z', t \mid \mathcal{X}) \, h \mathrm{d}y'
\end{aligned} \tag{B.1.1}$$

The existence of the density permits the definition of the *characteristic function* of the random variable $Y(t+h)$ as its Fourier transform:

$$\psi(\nu, t+h \mid y', t, \mathcal{X}) := \mathsf{E}\left[e^{i\nu(Y(t+h) - Y(t))} \mid Z(t+h) = z, Y(t) = y', Z(t) = z', \mathcal{X} \right].$$

The density can accordingly be expressed as the Fourier transform of the characteristic function: generally, the Fourier transform

$$\hat{f}(u) = \mathcal{F}\{f(y)\} := \int f(y) e^{-iu^\top y} \, \mathrm{d}y, \tag{B.1.2}$$

hence

$$p(y, t+h \mid z, t+h, y', z', t, \mathcal{X}) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\nu(y-y')} \psi(\nu, t+h \mid y', t, \mathcal{X}) \, \mathrm{d}\nu.$$

Within this Fourier transform, we expand the argument around $\nu = 0$:

$$\begin{aligned}
&\psi(\nu, t+h \mid y', t, \mathcal{X}) = \\
&\quad \sum_{n=0}^{\infty} \frac{(i\nu)^n}{n!} \mathsf{E}[(Y(t+h) - Y(t))^n \mid Z(t+h) = z, Y(t) = y', Z(t) = z', \mathcal{X}].
\end{aligned} \tag{B.1.3}$$

This allows us to utilize the identity (holding under the integral)

$$\partial_y^{(n)}\delta(y-y') = \frac{1}{2\pi}(-i\nu)^n \int_{\mathbb{R}} e^{-i\nu(y-y')}\,\mathrm{d}\nu, \tag{B.1.4}$$

with $\partial_y^{(0)}\delta(y-y') = \delta(y-y')$. Combining both expressions (B.1.3) and (B.1.4) in Eq. (B.1.1) yields

$$\begin{aligned}
&p(y,z,t+h \mid \mathcal{X}) \\
&= \int_{\mathcal{Y}} p(y,t+h \mid z,t+h,y',z,t,\mathcal{X})p(y',z,t \mid X)\,\mathrm{d}y' \\
&\qquad + h\cdot\sum_{z'}\int_{\mathcal{Y}} p(y,t+h \mid z,t+h,y',z',t,\mathcal{X})\Lambda_{z'z}^{y'}(t)p(y',z',t \mid \mathcal{X})\,\mathrm{d}y' \\
&= \int_{\mathcal{Y}}\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\partial_y^{(n)}\delta(y-y')\,\mathsf{E}[(Y(t+h)-Y(t))^n \mid Z(t+h)=z, \\
&\qquad\qquad\qquad\qquad Y(t)=y', Z(t)=z,\mathcal{X}]p(y',z,t \mid \mathcal{X})\,\mathrm{d}y' \\
&\qquad + h\cdot\sum_{z'}\int_{\mathcal{Y}}\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\partial_y^{(n)}\delta(y-y')\,\mathsf{E}[(Y(t+h)-Y(t))^n \mid Z(t+h)=z, \\
&\qquad\qquad\qquad\qquad Y(t)=y', Z(t)=z',\mathcal{X}]\Lambda_{z'z}^{y'}(t)p(y',z',t \mid \mathcal{X})\,\mathrm{d}y' \\
&= \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\partial_y^{(n)}\,\mathsf{E}[(Y(t+h)-Y(t))^n \mid Z(t+h)=z, \\
&\qquad\qquad\qquad\qquad Y(t)=y, Z(t)=z,\mathcal{X}]p(y,z,t \mid \mathcal{X}) \\
&\qquad + h\cdot\sum_{z'}\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\partial_y^{(n)}\,\mathsf{E}[(Y(t+h)-Y(t))^n \mid Z(t+h)=z, \\
&\qquad\qquad\qquad\qquad Y(t)=y, Z(t)=z',\mathcal{X}]\Lambda_{z'z}^{y}(t)p(y,z',t \mid \mathcal{X}).
\end{aligned}$$

As Y is an SSDE

$$\mathrm{d}Y(t) = f(Y(t),Z(t),t)\,\mathrm{d}t + Q(Y(t),Z(t),t)\,\mathrm{d}W(t),$$

and the discrete process remaining constant in a small time interval, $Z_{[t,t+h]} = z$, this SSDE can be treated as a conventional, Z-independent Itô SDE. For small h, we can utilize the usual Euler-Maruyama approximation,

$$Y(t+h) \mid Z(t+h)=z, Z(t)=z, Y(t)=y \sim \mathcal{N}(y+f(y,z,t)h, D(y,z,t)h).$$

With this we notice that in the second term above, only $n = 0$ contributes, as all other terms are at least of order $o(h)$. We accordingly find

$$\begin{aligned}
&p(y, z, t + h \mid \mathcal{X}) \\
&= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \partial_y^{(n)} \mathsf{E}[(Y(t+h) - Y(t))^n | Z(t+h) = z, \\
&\qquad Y(t) = y, Z(t) = z, \mathcal{X}] p(y, z, t \mid \mathcal{X}) + h \cdot \sum_{z'} \Lambda^y_{z'z}(t) p(y, z', t \mid \mathcal{X}) \\
&= p(y, z, t \mid \mathcal{X}) \\
&\quad + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \partial_y^{(n)} \mathsf{E}[(Y(t+h) - Y(t))^n | Z(t+h) = z, \\
&\qquad Y(t) = y, Z(t) = z, \mathcal{X}] p(y, z, t \mid \mathcal{X}) + h \cdot \sum_{z'} \Lambda^y_{z'z}(t) p(y, z', t \mid \mathcal{X})
\end{aligned}$$

Substracting $p(y, z, t \mid \mathcal{X})$ from both sides, dividing by h and taking the limit $h \to 0$ yields

$$\begin{aligned}
\partial_t p(y, z, t \mid \mathcal{X}) &= \lim_{h \to 0} \frac{p(y, z, t + h \mid \mathcal{X}) - p(y, z, t \mid \mathcal{X})}{h} \\
&= \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \partial_y^{(n)} \{\Gamma_{nyz} p(y, z, t \mid \mathcal{X})\} + \sum_{z'} \Lambda^y_{z'z}(t) p(y, z', t \mid \mathcal{X}) \qquad \text{(B.1.5)}
\end{aligned}$$

with

$$\begin{aligned}
\Gamma_{nyz} &= \lim_{h \to 0} \frac{1}{h} \mathsf{E}[(Y(t+h) - Y(t))^n \mid Z(t+h) = z, Y(t) = y, Z(t) = z, \mathcal{X}] \\
\Lambda^y_{z'z} &= \lim_{h \to 0} \frac{1}{h} p(z, t + h \mid z', y', t, \mathcal{X}) - \delta_{z'z}.
\end{aligned} \qquad \text{(B.1.6)}$$

As discussed above, when conditioned on the same mode z, Y can be treated as a conventional SDE; we can hence evaluate the conditional moments Γ_{nyz} in closed form. We have

$$\begin{aligned}
&\mathsf{E}[(Y(t+h) - Y(t))^n \mid Z(t+h) = z, Y(t) = y, Z(t) = z, \mathcal{X}] \\
&\qquad = \int (y' - y)^n \mathcal{N}(z' \mid y + f(y, z, t)h, D(y, z, t)h) \, \mathrm{d}y'
\end{aligned}$$

and the first two conditional moments are the usual Gaussian moments

$$\Gamma_{nyz} = \begin{cases} f(y, z, t) & \text{if } n = 1 \\ \frac{1}{2} Q(y, z, t) Q^\top(y, z, t) = \frac{1}{2} D(y, z, t) & \text{if } n = 2. \end{cases}$$

As shown in [102], if $\Gamma_{nyz} = 0$ for some even n, $\Gamma_{nyz} = 0 \; \forall n \geq 0$. It is straightforward to show that, e.g., $\Gamma_{nyz} = 0$ for $n = 4$, so all other conditional moments vanish. Hence, the above reduces to the HME

$$\partial_t p(y, z, t \mid \mathcal{X}) = \mathcal{L}_t^\dagger p(y, z, t \mid \mathcal{X})$$

with $\mathcal{L}_t^\dagger = \mathcal{F}_t^\dagger + \mathcal{T}_t^\dagger$ as

$$\begin{aligned}\mathcal{F}_t^\dagger p(y,z,t \mid \mathcal{X}) &= -\sum_{i=1}^n \partial_{y_i} \left\{ f_i(y,z,t) p(y,z,t \mid \mathcal{X}) \right\} \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \partial_{y_i} \partial_{y_j} \{ D_{ij}(y,z,t) p(y,z,t \mid \mathcal{X}) \}, \\ \mathcal{T}_t^\dagger p(y,z,t \mid \mathcal{X}) &= \sum_{z' \in \mathcal{Z} \setminus z} \Lambda(z',z,t) p(y,z',t \mid \mathcal{X}) - \Lambda(z,t) p(y,z,t \mid \mathcal{X}).\end{aligned}$$

Starting off the derivation with the Chapman-Kolmogorov equation in the other time direction,

$$p(\mathcal{X} \mid y,z,t-h) = \sum_{z' \in \mathcal{Z}} \int p(y',z',t \mid y,z,t-h) p(\mathcal{X} \mid y',z',t) \,\mathrm{d}y',$$

one finds another PDE for the density $p(\mathcal{X} \mid y,z,t)$. This yields the backward equation

$$\partial_t p(\mathcal{X} \mid y,z,t) = -\mathcal{L}_t p(\mathcal{X} \mid y,z,t),$$

with the the generator $\mathcal{L} = \mathcal{F} + \mathcal{T}$:

$$\begin{aligned}\mathcal{F} p(\mathcal{X} \mid y,z,t) &= \sum_{i=1}^n f_i(y,z,t) \partial_{y_i} p(\mathcal{X} \mid y,z,t) \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij}(y,z,t) \partial_{y_i} \partial_{y_j} p(\mathcal{X} \mid y,z,t), \\ \mathcal{T} p(\mathcal{X} \mid y,z,t) &= \sum_{z' \in \mathcal{Z} \setminus z} \Lambda(z,z',t) p(\mathcal{X} \mid y,z,t) - \Lambda(z,t) p(\mathcal{X} \mid y,z,t).\end{aligned}$$

B.1.2 *Derivation of the posterior hybrid master equation*

As shown in the main text, using $k = \max\{k' \in \mathbb{N} : t_{k'} \leq t\}$, the (exact) smoothing density $p(y,z,t \mid x_{[0,T]})$, obeys

$$p(y,z,t \mid x_{[0,T]}) = C^{-1}(t) \alpha(y,z,t) \beta(y,z,t), \tag{B.1.7}$$

with the (forward) filtering density $\alpha(y,z,t) := p(y,z,t \mid x_{[0,t_k]})$, the backward function $\beta(y,z,t) := p(x_{[t_{k+1},T]} \mid y,z,t)$, and the normalizer $C(t)$.

With the dynamics for α and β provided in the main text, we compute now the dynamics of Eq. (B.1.7). For brevity, define $\gamma(y,z,t) := p(y,z,t \mid x_{[0,T]}) = C^{-1}(t)\alpha(y,z,t)\beta(y,z,t)$. We have, as given in the main text,

$$\partial_t \gamma(y,z,t) = C^{-1}(t) \alpha(y,z,t) \partial_t \beta(y,z,t) + C^{-1}(t) \beta(y,z,t) \partial_t \alpha(y,z,t). \tag{B.1.8}$$

The dynamics of the filtering distribution

$$\begin{aligned}\partial_t\alpha(y,z,t) = &-\sum_{i=1}^{n}\partial_{y_i}\left\{f_i(y,z,t)\alpha(y,z,t)\right\}\\ &+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_{y_i}\partial_{y_j}\left\{D_{ij}(y,z,t)\alpha(y,z,t)\right\}+\sum_{z'\in\mathcal{Z}}\Lambda(z',z,t)\alpha(y,z',t).\end{aligned}$$

Similarly,

$$\begin{aligned}\partial_t\beta(y,z,t) = &-\sum_{i=1}^{n}f_i(y,z,t)\partial_{y_i}\beta(y,z,t)\\ &-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}D_{ij}(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t)-\sum_{z'\in\mathcal{Z}}\Lambda(z,z',t)\beta(y,z',t).\end{aligned}$$

Inserting these dynamics into Eq. (B.1.8) and using $C^{-1}(t)\alpha(y,z,t)=\frac{\gamma(y,z,t)}{\beta(y,z,t)}$ yields

$$\begin{aligned}&\partial_t\gamma(y,z,t)\\ &=\frac{\gamma(y,z,t)}{\beta(y,z,t)}\left(-\sum_{i=1}^{n}f_i(y,z,t)\partial_{y_i}\beta(y,z,t)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}D_{ij}(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t)\right.\\ &\quad\left.-\sum_{z'\in\mathcal{Z}}\Lambda(z,z',t)\beta(y,z',t)\right)\\ &\quad+\beta(y,z,t)\left(-\sum_{i=1}^{n}\partial_{y_i}\left\{f_i(y,z,t)\frac{\gamma(y,z,t)}{\beta(y,z,t)}\right\}\right.\\ &\quad\left.+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_{y_i}\partial_{y_j}\left\{D_{ij}(y,z,t)\frac{\gamma(y,z,t)}{\beta(y,z,t)}\right\}+\sum_{z'\in\mathcal{Z}}\Lambda(z',z,t)\frac{\gamma(y,z',t)}{\beta(y,z',t)}\right).\end{aligned}\tag{B.1.9}$$

Now, compute the occuring first and second order derivatives:

$$\begin{aligned}\partial_{y_i}\left\{f_i(y,z,t)\frac{\gamma(y,z,t)}{\beta(y,z,t)}\right\}&=\frac{\left\{\partial_{y_i}f_i(y,z,t)\gamma(y,z,t)+f_i(y,z,t)\partial_{y_i}\gamma(y,z,t)\right\}\beta(y,z,t)}{\beta(y,z,t)^2}\\ &\quad-\beta(y,z,t)^{-2}f_i(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\ &=\beta(y,z,t)^{-1}\left\{\partial_{y_i}f_i(y,z,t)\gamma(y,z,t)+f_i(y,z,t)\partial_{y_i}\gamma(y,z,t)\right\}\\ &\quad-\beta(y,z,t)^{-2}f_i(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t),\end{aligned}$$

and

$$\begin{aligned}&\partial_{y_i}\partial_{y_j}\left\{\frac{D_{ij}(y,z,t)\gamma(y,z,t)}{\beta(y,z,t)}\right\}\\ =&\partial_{y_i}\left\{\gamma(y,z,t)\beta(y,z,t)^{-1}\partial_{y_j}D_{ij}(y,z,t)+D_{ij}(y,z,t)\beta(y,z,t)^{-1}\partial_{y_j}\gamma(y,z,t)\right.\\ &\left.-D_{ij}(y,z,t)\gamma(y,z,t)\beta(y,z,t)^{-2}\partial_{y_j}\beta(y,z,t)\right\}\end{aligned}$$

$$\begin{aligned}
&= \big(\gamma(y,z,t)\partial_{y_i}\partial_{y_j}D_{ij}(y,z,t) + \partial_{y_i}D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)\big)\,\beta(y,z,t)^{-1}\\
&\quad - \gamma(y,z,t)\beta(y,z,t)^{-2}\partial_{y_j}D_{ij}(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\quad + \partial_{y_j}D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\beta(y,z,t)^{-1}\\
&\quad + D_{ij}(y,z,t)\,\big(\beta(y,z,t)^{-1}\partial_{y_i}\partial_{y_j}\gamma(y,z,t) - \beta(y,z,t)^{-2}\partial_{y_j}\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\big)\\
&\quad - \partial_{y_i}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\beta(y,z,t)^{-2}\\
&\quad - D_{ij}(y,z,t)\,\big(\partial_{y_i}\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\beta(y,z,t)^{-2}\\
&\qquad\quad + \gamma(y,z,t)[\partial_{y_i}\partial_{y_j}\beta(y,z,t)\beta(y,z,t)^{-2}\\
&\qquad\quad -2\partial_{y_j}\beta(y,z,t)\beta(y,z,t)^{-3}\partial_{y_i}\beta(y,z,t)]\big)\,.
\end{aligned}$$

Collecting terms in $\beta(y,z,t)^{-1}$, $\beta(y,z,t)^{-2}$ and $\beta(y,z,t)^{-3}$ structures this expression:

$$\begin{aligned}
&\partial_{y_i}\partial_{y_j}\left\{\frac{D_{ij}(y,z,t)\gamma(y,z,t)}{\beta(y,z,t)}\right\}\\
&\quad=\beta(y,z,t)^{-1}\,\big\{\partial_{y_i}\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t) + \partial_{y_j}D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\\
&\qquad\qquad +\partial_{y_i}D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t) + D_{ij}(y,z,t)\partial_{y_i}\partial_{y_j}\gamma(y,z,t)\big\}\\
&\qquad - \beta(y,z,t)^{-2}\,\big\{\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad + \partial_{y_i}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad + D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad + D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad +D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t)\big\}\\
&\qquad + \beta(y,z,t)^{-3}\,\big\{2D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\partial_{y_j}\beta(y,z,t)\big\}\,.
\end{aligned}$$

Using the terms in Eq. (B.1.9), it then follows that

$$\begin{aligned}
&\partial_t\gamma(y,z,t)\\
&= -\sum_{i=1}^{n} f_i(y,z,t)\gamma(y,z,t)\beta(y,z,t)^{-1}\partial_{y_i}\beta(y,z,t)\\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}(y,z,t)\gamma(y,z,t)\beta(y,z,t)^{-1}\partial_{y_i}\partial_{y_j}\beta(y,z,t)\\
&\quad -\sum_{i=1}^{n}\beta(y,z,t)^{-1}\,\{[\partial_{y_i}f_i(y,z,t)\gamma(y,z,t) + f_i(y,z,t)\partial_{y_i}\gamma(y,z,t)]\beta(y,z,t)\\
&\qquad\qquad -f_i(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\}\\
&\quad +\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big[\partial_{y_i}\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t) + \partial_{y_j}D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\\
&\qquad\qquad + \partial_{y_i}D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t) + D_{ij}(y,z,t)\partial_{y_i}\partial_{y_j}\gamma(y,z,t)\\
&\qquad\quad - \beta(y,z,t)^{-1}\,\big\{\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad + \partial_{y_i}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad + D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad + D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad +D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t)\big\}\\
&\qquad\quad + \beta(y,z,t)^{-2}\big\{2D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\partial_{y_j}\beta(y,z,t)\big\}\big]\\
&+\sum_{z'\in\mathcal{Z}}\left(\Lambda(z',z,t)\frac{\beta(y,z,t)}{\beta(y,z',t)}\gamma(y,z',t)-\Lambda(z,z',t)\frac{\beta(y,z',t)}{\beta(y,z,t)}\gamma(y,z,t)\right)\\
=&-\sum_{i=1}^{n}\partial_{y_i}f_i(y,z,t)\gamma(y,z,t)+f_i(y,z,t)\partial_{y_i}\gamma(y,z,t)\\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_{y_i}\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t)+\partial_{y_j}D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\\
&\qquad\quad +\partial_{y_i}D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)+D_{ij}(y,z,t)\partial_{y_i}\partial_{y_j}\gamma(y,z,t)\\
&-\frac{\beta(y,z,t)^{-1}}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big\{\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad\qquad +\partial_{y_i}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad\qquad +2D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad\qquad +D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad\qquad +D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\big\}\\
&+\beta(y,z,t)^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\partial_{y_j}\beta(y,z,t)\\
&+\sum_{z'\in\mathcal{Z}\backslash z}\left(\Lambda(z',z,t)\frac{\beta(y,z,t)}{\beta(y,z',t)}\gamma(y,z',t)-\Lambda(z,z',t)\frac{\beta(y,z',t)}{\beta(y,z,t)}\gamma(y,z,t)\right)\\
=&-\sum_{i=1}^{n}\big\{\partial_{y_i}f_i(y,z,t)\gamma(y,z,t)+f_i(y,z,t)\partial_{y_i}\gamma(y,z,t)\\
&\qquad\quad +\frac{\beta(y,z,t)^{-1}}{2}\sum_{j=1}^{n}\big[\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad\qquad +\partial_{y_i}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad\qquad +2D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t)\\
&\qquad\qquad\qquad +D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\\
&\qquad\qquad\qquad +D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\big]\\
&\qquad\quad -\beta(y,z,t)^{-2}\sum_{j=1}^{n}D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\partial_{y_j}\beta(y,z,t)\Bigg\}\\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_{y_i}\big\{\partial_{y_j}D_{ij}(y,z,t)\gamma(y,z,t)+D_{ij}(y,z,t)\partial_{y_j}\gamma(y,z,t)\big\}\\
&+\sum_{z'\in\mathcal{Z}}\Lambda(z',z,t)\frac{\beta(y,z,t)}{\beta(y,z',t)}\gamma(y,z',t)\\
=&-\sum_{i=1}^{n}\big\{\partial_{y_i}\big\{f_i(y,z,t)\gamma(y,z,t)\big\}
\end{aligned}
$$

$$
\begin{aligned}
&+ \beta(y,z,t)^{-1} \sum_{j=1}^{n} \big[\partial_{y_i} D_{ij}(y,z,t) \gamma(y,z,t) \partial_{y_j} \beta(y,z,t) \\
&\qquad\qquad + D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\partial_{y_j}\beta(y,z,t) \\
&\qquad\qquad + D_{ij}(y,z,t)\partial_{y_i}\gamma(y,z,t)\partial_{y_j}\beta(y,z,t)\big] \\
&-\beta(y,z,t)^{-2} \sum_{j=1}^{n} D_{ij}(y,z,t)\gamma(y,z,t)\partial_{y_i}\beta(y,z,t)\partial_{y_j}\beta(y,z,t) \Bigg\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j} \left\{ D_{ij}(y,z,t)\gamma(y,z,t) \right\} + \sum_{z' \in \mathcal{Z}} \Lambda(z',z,t) \frac{\beta(y,z,t)}{\beta(y,z',t)} \gamma(y,z',t).
\end{aligned}
$$

Next, make use of the product rule to collect terms:

$$
\begin{aligned}
&\partial_t \gamma(y,z,t) \\
&= - \sum_{i=1}^{n} \partial_{y_i} \left\{ f_i(y,z,t)\gamma(y,z,t) + \sum_{j=1}^{n} D_{ij}(y,z,t)\partial_{y_j}\beta(y,z,t)\beta(y,z,t)^{-1}\gamma(y,z,t) \right\} \\
&\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j} \left\{ D_{ij}(y,z,t)\gamma(y,z,t) \right\} + \sum_{z' \in \mathcal{Z}} \Lambda(z',z,t) \frac{\beta(y,z,t)}{\beta(y,z',t)} \gamma(y,z',t) \\
&= - \sum_{i=1}^{n} \partial_{y_i} \left\{ \left(f_i(y,z,t) + \sum_{j=1}^{n} D_{ij}(y,z,t)\partial_{y_j} \ln \beta(y,z,t) \right) \gamma(y,z,t) \right\} \\
&\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j} \left\{ D_{ij}(y,z,t)\gamma(y,z,t) \right\} + \sum_{z' \in \mathcal{Z}} \Lambda(z',z,t) \frac{\beta(y,z,t)}{\beta(y,z',t)} \gamma(y,z',t).
\end{aligned}
$$

Finally, this gives the result as

$$
\begin{aligned}
\partial_t p(y,z,t \mid x_{[0,T]}) =& \tilde{\mathcal{L}}_t^{\dagger} p(y,z,t \mid x_{[0,T]}) \\
=& - \sum_{i=1}^{n} \partial_{y_i} \left\{ \tilde{f}_i(y,z,t) p(y,z,t \mid x_{[0,T]}) \right\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{y_i}\partial_{y_j} \left\{ \tilde{D}_{ij}(y,z,t) p(y,z,t \mid x_{[0,T]}) \right\} \\
&+ \sum_{z' \in \mathcal{Z}} \tilde{\Lambda}(y,z',z,t) p(y,z',t \mid x_{[0,T]}),
\end{aligned}
$$

with the posterior drift

$$
\tilde{f}_i(y,z,t) = f_i(y,z,t) + \sum_{j=1}^{n} D_{ij}(y,z,t)\partial_{y_j} \ln \beta(y,z,t),
$$

the posterior dispersion

$$
\tilde{D}_{ij}(y,z,t) = D_{ij}(y,z,t),
$$

and the posterior rate

$$
\tilde{\Lambda}(y,z',z,t) = \Lambda(z',z,t) \frac{\beta(y,z,t)}{\beta(y,z',t)}.
$$

B.2 SAMPLING THE CONDITIONAL DIFFUSION PROCESS

B.2.1 *Derivation of the backward continuous-discrete Kalman filter*

As discussed in the main text, we utilize the method of characteristics to explicitly solve the backward Kalman-type filter. The idea of this method is the following [110]: we aim to reparameterize the equation in such a way from the usual state-space and time variables y, t to some other path parameters γ and the initial values of the problem y_0, that along the curves parameterized by γ, the PDE to be solved reduces to an ODE system.

The backward distribution $p(x_N \mid y, t)$ between the N-th and $(N-1)$-th observation is given by the Kolmogorov backward equation (KBE)

$$\partial_t p(x_N \mid y, t) = -\mathcal{L}_t p(x_N \mid y, t), \tag{B.2.10}$$

where we assume an end-point condition

$$p(x_N \mid y, T) = \mathcal{N}(x_N \mid Fy, \Sigma).$$

In the linear case considered here,

$$\mathcal{L}_t(\cdot) = (\partial_y(\cdot))^\top (A(t)y + b(t)) + \frac{1}{2}\operatorname{tr}\left(D\partial_y\partial_y^\top(\cdot)\right).$$

Hence,

$$\partial_t p(x_N \mid y, t) = -\left(\partial_y p(x_N \mid y, t)\right)^\top (A(t)y + b(t)) - \frac{1}{2}\operatorname{tr}\left(D\partial_y\partial_y^\top p(x_N \mid y, t)\right). \tag{B.2.11}$$

Component-wise, this is

$$\begin{aligned}\partial_t p(x_N \mid y, t) = -\sum_{k,l} \partial_{y_k} p(x_N \mid y, t) A_{kl}(t) y_l - \sum_k \partial_{y_k} p(x_N \mid y, t) b_k(t)\\ - \frac{1}{2}\sum_{k,l} \partial_{y_k}\partial_{y_l} p(x_N \mid y, t) D_{kl}.\end{aligned} \tag{B.2.12}$$

We utilize again the Fourier transform Eq. (B.1.2), which obeys the following identities:

$$\begin{aligned}\mathcal{F}\{f(Fy)\} &= \frac{1}{|F|}\hat{f}(F^{-\top}u),\\ \mathcal{F}\{\partial_y f(y)\} &= iu\hat{f}(u),\\ \mathcal{F}\{y\partial_y f(y)\} &= i\partial_u \hat{f}(u).\end{aligned} \tag{B.2.13}$$

With this, the Fourier transform $\partial_t \hat{p}(x_N \mid u,t)$ is obtained as

$$\begin{aligned}\partial_t \hat{p}(x_N \mid u,t) &= -\sum_{k,l} i\partial_{u_l} \mathcal{F}\{\partial_{y_k} p(x_N \mid y,t) A_{kl}(t)\} - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) \\ &\qquad - \frac{1}{2}\sum_{k,l} u_k u_l \hat{p}(x_N \mid u,t) D_{kl} \\ &= -\sum_{k,l} i\partial_{u_l}\{i u_k \hat{p}(x_N \mid u,t) A_{kl}(t)\} - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) \\ &\qquad - \frac{1}{2}\sum_{k,l} u_k u_l \hat{p}(x_N \mid u,t) D_{kl} \\ &= \sum_{k,l} \partial_{u_l} u_k \hat{p}(x_N \mid u,t) A_{kl}(t) + \sum_{k,l} u_k \partial_{u_l} \hat{p}(x_N \mid u,t) A_{kl}(t) \\ &\quad - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) - \frac{1}{2}\sum_{k,l} i u_k i u_l \hat{p}(x_N \mid u,t) D_{kl} \\ &= \sum_{k,l} \mathbb{1}(k=l) \hat{p}(x_N \mid u,t) A_{kl}(t) + \sum_{k,l} u_k \partial_{u_l} \hat{p}(x_N \mid u,t) A_{kl}(t) \\ &\quad - \sum_k i u_k \hat{p}(x_N \mid u,t) b_k(t) + \frac{1}{2}\sum_{k,l} u_k u_l \hat{p}(x_N \mid u,t) D_{kl}.\end{aligned}$$

In vector notation, this reads

$$\begin{aligned}\partial_t \hat{p}(x_N \mid u,t) &= \operatorname{tr}(A(t)) \hat{p}(x_N \mid u,t) + (\partial_u \hat{p}(x_N \mid u,t))^\top A^\top(t) u \\ &\qquad - i u^\top b(t) \hat{p}(x_N \mid u,t) + \frac{1}{2} u^\top D u \hat{p}(x_N \mid u,t). \qquad \text{(B.2.14)}\end{aligned}$$

Note that while we retain a Z-dependency of D, $D = D(z)$, this function is constant between switches of Z; as we solve the backward filter piece-wise between jumps in the given sample $z_{[0,T]}$, this dependency is irrevelant at this point. Now, we define the characteristic curve as

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t) = -A^\top(t) u(t), \qquad \text{(B.2.15)}$$

with end-point condition $u(T) = u_T$. The formal solution to this equation is expressible via the transition function $\Phi(t,T)$ [20]

$$u(t) = \Phi^\top(t,T) u_T, \qquad \text{(B.2.16)}$$

where Φ is defined by the relations

$$\begin{aligned}\partial_T \Phi(t,T) &= -A(T)\Psi(t,T), & \partial_t \Psi(t,T) &= \Psi(t,T)A(t), \\ \Psi(t,T) &= \Psi(t',T)\Psi(t,t'), & \Psi(T,t) &= \Psi(t,T)^{-1},\end{aligned}$$

together with $\Psi(t,t) = \mathbb{1}$. Hence, we have

$$\begin{aligned}u_T &= \Phi^{-\top}(t,T) u(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,T) &= -\Phi(t,T)A(t),\end{aligned} \qquad \text{(B.2.17)}$$

with which the total derivative of $\hat{p}(x_N \mid u, t)$ at $u = u(t)$ is found as

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{p}(x_N \mid u(t), t) = (\partial_u \hat{p}(x_N \mid u(t), t))^\top \frac{\mathrm{d}}{\mathrm{d}t}u(t) + \partial_t \hat{p}(x_N \mid u(t), t). \tag{B.2.18}$$

Inserting the backward quantity $\partial_t \hat{p}(x_N \mid u(t), t)$ and the characteristic ODE (B.2.15), we find

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{p}(x_N \mid u(t), t) = \left(\mathrm{tr}(A(t)) - iu^\top(t)b(t) + \frac{1}{2}u^\top(t)Du(t)\right)\hat{p}(x_N \mid u(t), t), \tag{B.2.19}$$

yielding the solution

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) = \exp\bigg\{&-\int_t^T \mathrm{tr}(A(s))\,\mathrm{d}s - i\left(-\int_t^T u^\top(s)b(s)\,\mathrm{d}s\right)\\ &-\frac{1}{2}\int_t^T u^\top(s)Du(s)\,\mathrm{d}s\bigg\}\,\hat{p}(x_N \mid u(T), T).\end{aligned} \tag{B.2.20}$$

To obtain the end-point condition in Fourier space, $\hat{p}(x_N \mid u(T), T) = \hat{p}(x_N \mid u_T, T)$, we compute

$$\begin{aligned}\mathcal{F}\{p(x_N \mid y, T)\} &= \mathcal{F}\{\mathcal{N}(x_N \mid Fy, \Sigma)\}\\ &= \mathcal{F}\{\mathcal{N}(Fy \mid x_N, \Sigma)\}\\ &= \frac{1}{|F|}\exp\left\{-i\left(F^{-\top}u\right)^\top x_N - \frac{1}{2}\left(F^{-\top}u\right)^\top \Sigma F^{-\top}u\right\},\end{aligned} \tag{B.2.21}$$

where we assume F to be quadratic and invertible. Hence,

$$\hat{p}(x_N \mid u(T), T) = \frac{1}{|F|}\exp\left\{-i\left(F^{-\top}u_T\right)^\top x_N - \frac{1}{2}\left(F^{-\top}u_T\right)^\top \Sigma F^{-\top}u_T\right\}, \tag{B.2.22}$$

and we have

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) = \frac{1}{|F|}\exp\bigg\{&-\int_t^T \mathrm{tr}(A(s))\,\mathrm{d}s\\ &- i\left(\left(F^{-\top}u_T\right)^\top x_N - \int_t^T u^\top(s)b(s)\,\mathrm{d}s\right)\\ &-\frac{1}{2}\left(\left(F^{-\top}u_T\right)^\top \Sigma F^{-\top}u_T + \int_t^T u^\top(s)Du(s)\,\mathrm{d}s\right)\bigg\}.\end{aligned} \tag{B.2.23}$$

Inserting the formal solution Eq. (B.2.16) gives

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) &= \frac{1}{|F|}\exp\bigg\{-\int_t^T \mathrm{tr}(A(s))\,\mathrm{d}s\\ &\quad - i\left(\left(F^{-\top}u_T\right)^\top x_N - u_T^\top \int_t^T \Phi(s,T)b(s)\,\mathrm{d}s\right)\\ &\quad -\frac{1}{2}\left(\left(F^{-\top}u_T\right)^\top \Sigma F^{-\top}u_T + u_T^\top \int_t^T \Phi(s,T)D\Phi^\top(s,T)\,\mathrm{d}s\,u_T\right)\bigg\}\\ &= \frac{1}{|F|}\exp\bigg\{-\int_t^T \mathrm{tr}(A(s))\,\mathrm{d}s - iu_T^\top F^{-1}\left(x_N - F\int_t^T \Phi(s,T)b(s)\,\mathrm{d}s\right)\\ &\quad -\frac{1}{2}u_T^\top F^{-1}\left(\Sigma + F\int_t^T \Phi(s,T)D\Phi^\top(s,T)\,\mathrm{d}s\,F^\top\right)F^{-\top}u_T\bigg\}.\end{aligned} \tag{B.2.24}$$

Using the transpose of Eq. (B.2.16), we find

$$\begin{aligned}\hat{p}(x_N \mid u(t), t) = \frac{1}{|F|} \exp\Bigg\{ &- \int_t^T \operatorname{tr}(A(s))\, \mathrm{d}s \\ &- i u^\top(t) \Phi^{-1}(t,T) F^{-1} \left(x_N - F \int_t^T \Phi(s,T) b(s)\, \mathrm{d}s \right) \\ -\frac{1}{2} u^\top(t) \Phi^{-1}(t,T) F^{-1} &\left(\Sigma + F \int_t^T \Phi(s,T) D \Phi^\top(s,T)\, \mathrm{d}s\, F^\top \right) F^{-\top} \Phi^{-\top}(t,T) u(t) \Bigg\} \\ = \frac{1}{|F|} \exp\Bigg\{ &- \int_t^T \operatorname{tr}(A(s))\, \mathrm{d}s \\ - i \left((F\Phi(t,T))^{-\top} u(t) \right)^\top &\left(x_N - F \int_t^T \Phi(s,T) b(s)\, \mathrm{d}s \right) \\ - \frac{1}{2} \left((F\Phi(t,T))^{-\top} u(t) \right)^\top & \\ \cdot \left(\Sigma + F \int_t^T \Phi(s,T) D \Phi^\top(s,T)\, \mathrm{d}s\, F^\top \right) &(F\Phi(t,T))^{-\top} u(t) \Bigg\}.\end{aligned}$$

Inverting the Fourier transform yields

$$\begin{aligned} p(x_N \mid y, t) &= \frac{|\Phi(t,T)|}{\exp\left(\int_t^T \operatorname{tr}(A(s))\, \mathrm{d}s \right)} \\ &\cdot \mathcal{N}\left(F\Phi(t,T) y \,\middle|\, x_N - F \int_t^T \Phi(s,T) b(s)\, \mathrm{d}s, \Sigma + F \int_t^T \Phi(s,T) D \Phi^\top(s,T)\, \mathrm{d}s\, F^\top \right) \\ &= \frac{|\Phi(t,T)|}{\exp\left(\int_t^T \operatorname{tr}(A(s))\, \mathrm{d}s \right)} \\ &\cdot \mathcal{N}\left(x_N \,\middle|\, F\Phi(t,T) y + F \int_t^T \Phi(s,T) b(s)\, \mathrm{d}s, \Sigma + F \int_t^T \Phi(s,T) D \Phi^\top(s,T)\, \mathrm{d}s\, F^\top \right). \end{aligned} \tag{B.2.25}$$

With Jacobi's formula, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} |\Phi(t,T)| &= |\Phi(t,T)| \operatorname{tr}\left(\Phi^{-1}(t,T) \frac{\mathrm{d}}{\mathrm{d}t} \Phi(t,T) \right) \\ &= |\Phi(t,T)| \operatorname{tr}\left(\Phi^{-1}(t,T) \left(-\Phi(t,T) A(t) \right) \right) \\ &= |\Phi(t,T)| \operatorname{tr}\left(-A(t) \right), \end{aligned} \tag{B.2.26}$$

the solution of which reads

$$|\Phi(t,T)| = \exp\left(- \int_t^T \operatorname{tr}\left(-A(s) \right)\, \mathrm{d}s \right) \Phi(T,T) \tag{B.2.27}$$

with $\Phi(T,T) = \mathbb{I}$. Consequently,

$$|\Phi(t,T)| = \exp\left(\int_t^T \operatorname{tr}\left(A(s) \right)\, \mathrm{d}s \right), \tag{B.2.28}$$

and

$$p(x_N \mid y, t) = \mathcal{N}\left(x_N \,\middle|\, F\Phi(t,T)y + F\int_t^T \Phi(s,T)b(s)\,\mathrm{d}s,\right.$$
$$\left. \Sigma + F\int_t^T \Phi(s,T)D\Phi^\top(s,T)\,\mathrm{d}s\,F^\top\right). \quad \text{(B.2.29)}$$

Defining

$$F(t) := F\Phi(t,T), \quad \text{(B.2.30)}$$
$$m(t) := F\int_t^T \Phi(s,T)b(s)\,\mathrm{d}s = \int_t^T F(s)b(s)\,\mathrm{d}s, \quad \text{(B.2.31)}$$
$$\Sigma(t) := \Sigma + F\int_t^T \Phi(s,T)D\Phi^\top(s,T)\,\mathrm{d}s\,F^\top = \Sigma + \int_t^T F(s)DF^\top(s)\,\mathrm{d}s, \quad \text{(B.2.32)}$$

we have obtained the solution of the KBE (B.2.12) along the characteristic curve parameterized via $F(t), m(t)$ and $\Sigma(t)$. By straightforward differentiation, we obtain ODEs for these quantities, effectively reducing the solution of the original PDE to the solution of these ODEs. We find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(t) &= F\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,T)\\ &= F\left(-\Phi(t,T)A(t)\right)\\ \iff \frac{\mathrm{d}}{\mathrm{d}t}F(t) &= -F(t)A(t),\end{aligned} \quad \text{(B.2.33)}$$

with end-point condition $F(T) = F$.

Utilizing Leibniz' integral rule, we further have

$$\frac{\mathrm{d}}{\mathrm{d}t}m(t) = -F(t)b(t), \quad \text{(B.2.34)}$$

with boundary condition $m(T) = 0$.

Finally,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) = -F(t)DF^\top(t), \quad \text{(B.2.35)}$$

with boundary condition $\Sigma(T) = \Sigma$.

In summary,

$$p(x_N \mid y, t) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right), \quad \text{(B.2.36)}$$

with

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F(t) &= -F(t)A(t) && \text{with} && F(T) = F,\\ \frac{\mathrm{d}}{\mathrm{d}t}m(t) &= -F(t)b(t) && \text{with} && m(T) = 0,\\ \frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) &= -F(t)DF^\top(t) && \text{with} && \Sigma(T) = \Sigma.\end{aligned} \quad \text{(B.2.37)}$$

B.2.1.1 *The non-invertible case*

Above it was assumed that F be invertible. The solution however holds for general matrices F, which can be shown by plugging in the solution $p(x_N \mid y, t) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right)$ with some generaly, non-invertable $F(t)$ into the KBE (B.2.12). This yields the PDE

$$\begin{aligned} \partial_t \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) = &-\sum_{k,l} \partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) A_{kl}(t) y_l \\ &- \sum_k \partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) b_k(t) \\ &- \frac{1}{2} \sum_{k,l} \partial_{y_k} \partial_{y_l} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) D_{kl}. \end{aligned} \tag{B.2.38}$$

We now compute one-by-one the partial derivatives

$$\begin{aligned} &\partial_t \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right), \\ &\partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right), \\ &\partial_{y_k} \partial_{y_l} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right). \end{aligned}$$

To this end, notice that

$$\begin{aligned} \partial_\theta \mathcal{N}(x \mid a, A) = \mathcal{N}(x \mid a, A) \Big(&-h^\top(\partial_\theta x) + h^\top(\partial_\theta a) \\ &-\frac{1}{2} \operatorname{tr}(A^{-1} \partial_\theta A) + \frac{1}{2} h^\top (\partial_\theta A) h\Big), \end{aligned} \tag{B.2.39}$$

with $h = A^{-1}(x - a)$. This yields

$$\begin{aligned} \partial_t \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) &= \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) \\ &\cdot \left[h^\top(\partial_t F(t) y + \partial_t m(t)) - \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}(t) \partial_t \Sigma(t)\right) + \frac{1}{2} h^\top \partial_t \Sigma h\right], \end{aligned} \tag{B.2.40}$$

$$\partial_{y_k} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) = \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) \left[h^\top F_{\cdot k}(t)\right], \tag{B.2.41}$$

$$\begin{aligned} \partial_{y_k} \partial_{y_l} \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) &= \mathcal{N}\left(x_N \mid F(t)y + m(t), \Sigma(t)\right) \\ &\cdot \left[h^\top F_{\cdot k}(t) h^\top F_{\cdot l}(t) - L_{\cdot l}^\top(t) \Sigma^{-1}(t) F_{\cdot k}(t)\right], \end{aligned} \tag{B.2.42}$$

with $h = \Sigma^{-1}(t)(x_N - F(t)y - m(t))$. Inserting these equations into the KBE (B.2.38) yields

$$\begin{aligned} &h^\top(\partial_t F(t) y + \partial_t m(t)) - \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}(t) \partial_t \Sigma(t)\right) + \frac{1}{2} h^\top \partial_t \Sigma h \\ &= -\sum_k h^\top F_{\cdot k,l}(t) A_{kl}(t) y_l - \sum_k h^\top F_{\cdot k}(t) b_k(t) \\ &\qquad - \frac{1}{2} \sum_{k,l} \left(h^\top F_{\cdot k}(t) h^\top F_{\cdot l}(t) - F_{\cdot l}^\top(t) \Sigma^{-1}(t) F_{\cdot k}(t)\right) D_{kl}, \end{aligned} \tag{B.2.43}$$

or, in vector notation,

$$\begin{aligned} &h^\top \left[\partial_t F(t) y + \partial_t m(t)\right] - \frac{1}{2} \operatorname{tr}\left\{\left(\Sigma^{-1}(t) - h h^\top\right) \partial_t \Sigma(t)\right\} \\ &= h^\top \left[-F(t) A(t) y - F(t) b(t)\right] - \frac{1}{2} \operatorname{tr}\left\{\left(\Sigma^{-1}(t) - h h^\top\right) \left[-F^\top(t) D F(t)\right]\right\} \end{aligned} \tag{B.2.44}$$

Comparing coefficients, we find

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}F(t) &= -F(t)A(t) && \text{with} \quad F(T) = F, \\ \frac{\mathrm{d}}{\mathrm{d}t}m(t) &= -F(t)b(t) && \text{with} \quad m(T) = 0, \\ \frac{\mathrm{d}}{\mathrm{d}t}\Sigma(t) &= -F(t)DF^\top(t) && \text{with} \quad \Sigma(T) = \Sigma, \end{aligned} \tag{B.2.45}$$

where we identify the end-point conditions via

$$\mathcal{N}(x_N \mid F(T)y + m(T), \Sigma(T)) = \mathcal{N}(x_N \mid Fy, \Sigma). \tag{B.2.46}$$

JUMP CONDITIONS Starting at the end point, consider the last observation X_{N-1} at time point t_{N-1}. As shown in the main paper,

$$\beta(y, t_{N-1}) = \beta(y, t_{N-1}^+)p(x_{N-1} \mid y, t_{N-1}), \tag{B.2.47}$$

where $\beta(y, t_{N-1}^+) := \lim_{h \searrow 0} \beta(y, t_{N-1} + h)$. Assuming Gaussian observation likelihoods, we have, due to the properties of Gaussians,

$$\begin{aligned} \beta(y, t_{N-1}) &= \mathcal{N}\left(x_N \mid F(t_{N-1}^+)y + m(t_{N-1}^+), \Sigma(t_{N-1}^+)\right) \mathcal{N}(x_{N-1} \mid y, \Sigma_x) && \text{(B.2.48)} \\ &= \mathcal{N}\left(x_{N-1}, x_N \mid F(t_{N-1})y + m(t_{N-1}), \Sigma(t_{N-1})\right), && \text{(B.2.49)} \end{aligned}$$

with

$$F(t_{N-1}) = \begin{pmatrix} \mathbb{1}_{n\times n} \\ F(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{2n\times n}, \tag{B.2.50}$$

$$m(t_{N-1}) = \begin{pmatrix} 0 \\ m(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{2n}, \tag{B.2.51}$$

$$M(t_{N-1}) = \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma(t_{N-1}^+) \end{pmatrix} \in \mathbb{R}^{2n\times 2n}, \tag{B.2.52}$$

where $0 \in \mathbb{R}^n$ and $\mathbb{1}_{n\times n}$ is the n-dimensional identity matrix.

B.2.1.2 *Information filter parameterization*

The derived backward filter has the property that its support increases with every incorporated observation, which is computationally disadvantageous. The contribution of the backward filter to the drift of the posterior SDE is however fixed in size:

$$\mathrm{d}Y(t) = (f(Y(t), t) + D(Z(t))\partial \ln \beta(Y(t), t))\, \mathrm{d}t + Q(Z(t))\mathrm{d}W(t).$$

From the above derivations, we know that

$$\begin{aligned} \partial_y \ln \beta(y, t) &= -\frac{1}{2}\partial_y (x(t) - F(t)y(t) - m(t))^\top \Sigma^{-1}(t)(x(t) - F(t)y(t) - m(t)) \\ &= F(t)^\top \Sigma^{-1}(t)(x(t) - m(t)) - F(t)^\top \Sigma^{-1}(t)F(t)y(t) && \text{(B.2.53)} \end{aligned}$$

Defining

$$\begin{aligned} \nu(t) &:= F(t)^\top \Sigma^{-1}(t)(x(t) - m(t)), \\ M(t) &:= F(t)^\top \Sigma^{-1}(t) F(t), \end{aligned} \tag{B.2.54}$$

we can compute straightforwardly the respective time derivatives, where the notation $\dot{f} = \frac{\mathrm{d}}{\mathrm{d}t} f$ is used for conciseness:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M(t) &= \dot{F}(t)^\top \Sigma^{-1}(t) F(t) + F(t)^\top \dot{\Sigma}^{-1}(t) F(t) + F(t)^\top \Sigma^{-1}(t) \dot{F}(t) \\ &= -A(t)^\top M(t) + F(t)^\top \dot{\Sigma}^{-1}(t) F(t) - M(t) A(t) \\ &= -A(t)^\top M(t) - F(t)^\top \Sigma^{-1}(t) \dot{\Sigma}(t) \Sigma^{-1}(t) F(t) - M(t) A(t) \\ &= -A(t)^\top M(t) + F(t)^\top \Sigma^{-1}(t) F(t) D F(t)^\top \Sigma^{-1}(t) F(t) - M(t) A(t) \\ &= -A(t)^\top M(t) + M(t) D M(t) - M(t) A(t), \end{aligned} \tag{B.2.55}$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \nu(t) &= \dot{F}(t)^\top \Sigma^{-1}(t)(x(t) - m(t)) + F(t)^\top \dot{\Sigma}^{-1}(t)(x(t) - m(t)) \\ &\qquad\qquad\qquad\qquad - F(t)^\top \Sigma^{-1}(t) \dot{m}(t) \\ &= -A(t)^\top \nu(t) + M(t) D \nu(t) + M(t) b(t). \end{aligned} \tag{B.2.56}$$

The jump conditions for the information filter given in the main text follow directly by comparing Eq. (3.3.32) and Eq. (B.2.54).

B.3 SAMPLING THE CONDITIONAL SWITCHING PROCESS

The following is along the lines of the thorough treatments for the conventional smoothing problem in [31] and [55].

B.3.1 *Backward dynamics*

First, starting from the Radon-Nikodym derivative Eq. (3.3.42), write

$$G(Y_{[t,T]}, Z_{[t,T]}) = \exp\{\phi(t)\}, \tag{B.3.57}$$

defining the stochastic process

$$\begin{aligned} \phi(t) := \int_t^T & f(Y(s), Z(s))^\top D(Z(s), s)^{-1} \mathrm{d}Y(s) \\ & - \frac{1}{2} \int_t^T f(Y(s), Z(s))^\top D(Z(s), s)^{-1} f(Y(s), Z(s)) \mathrm{d}s. \end{aligned} \tag{B.3.58}$$

For brevity, set

$$G_{[t,T]} := G(Y_{[t,T]}, Z_{[t,T]}).$$

It is immediate that $\phi(T) = 0$; hence

$$\phi(T) - \phi(t) = \int_t^T \mathrm{d}\phi(s) \iff \phi(t) = -\int_t^T \mathrm{d}\phi(s) \tag{B.3.59}$$

and the differential

$$\begin{aligned} \mathrm{d}\phi(t) &= -f(Y(t), Z(t))^\top D(Z(t), t)^{-1} \mathrm{d}Y(t) \\ &\qquad + \frac{1}{2} f(Y(t), Z(t))^\top D(Z(t), t)^{-1} f(Y(t), Z(t)) \mathrm{d}t. \end{aligned} \tag{B.3.60}$$

This is an Itô SDE evolving *forward* in time—meaning that future increments of the process $y(t)$ are independent of its past—with a *terminal* condition $\phi(T) = 0$.

Applying Itô's lemma to Eq. (B.3.57) with respect to the stochastic process $\phi(t)$ results in

$$\begin{aligned} \mathrm{d}G_{[t,T]} &= -G_{[t,T]} f(Y(t), Z(t))^\top D^{-1}(Z(t), t) \mathrm{d}Y(t) \\ &\qquad + G_{[t,T]} f(Y(t), Z(t)) D^{-1}(Z(t), t) f(Y(t), Z(t))^\top \mathrm{d}t \end{aligned} \tag{B.3.61}$$

as the quadratic term $\mathrm{d}Y(t)\mathrm{d}Y(t)^\top = D(Z(t), t)\mathrm{d}t$. This can be written in integral from as

$$\begin{aligned} G_{[t,T]} = 1 &+ \int_t^T G_{[s,T]} f(Y(s), Z(s))^\top D^{-1}(Z(s), s) \mathrm{d}Y(s) \\ &- \int_t^T G_{[s,T]} f(Y(s), Z(s))^\top D^{-1}(Z(s), s) f(Y(s), Z(s)) \mathrm{d}s. \end{aligned}$$

Now, recall that $v(z, t) = \mathsf{E}[G_{[t,T]} \mid Y_{[t,T]} = y_{[t,T]}, Z(t) = z]$. To emphasize that we are conditioning on the full path $y_{[t,T]}$ drawn in the previous Gibbs step, we write $y(t)$ instead of $Y(t)$ in the following. With this, we obtain

$$\begin{aligned} v(z, t) &= 1 + \int_t^T \mathsf{E}\big[G_{[s,T]} f(y(s), Z(s))^\top D^{-1}(Z(s), s) \mid y_{[t,T]}, Z(t) = z\big] \, \mathrm{d}y(s) \\ &\quad - \int_t^T \mathsf{E}\big[G_{[s,T]} f(y(s), Z(s))^\top D^{-1}(Z(s), s) f(y(s), Z(s)) \mid y_{[t,T]}, Z(t) = z\big] \, \mathrm{d}s \\ &= 1 + \int_t^T \sum_{z' \in \mathcal{Z}} \mathsf{E}\big[G_{[s,T]} f(y(s), z')^\top D^{-1}(z', s) \mid y_{[s,T]}, z'\big] \, p(z', t \mid z, t) \mathrm{d}y(s) \\ - \int_t^T &\sum_{z' \in \mathcal{Z}} \mathsf{E}\big[G_{[s,T]} f(y(s), z')^\top D^{-1}(z', s) f(y(s), z') \mid y_{[t,T]}, Z(s) = z'\big] \, p(z', s \mid z, t) \mathrm{d}s \\ &= 1 + \int_t^T \sum_{z' \in \mathcal{Z}} v(z', s) f(y(s), z')^\top D^{-1}(z', s) p(z', s \mid z, t) \mathrm{d}y(s) \\ &\quad - \int_t^T \sum_{z' \in \mathcal{Z}} v(z', s) f(y(s), z')^\top D^{-1}(z', s) f(y(s), z') p(z', s \mid z, t) \mathrm{d}s. \end{aligned}$$

To compute the differential, consider with some small $h > 0$

$$\begin{aligned}\Delta v(z,t) &= v(z,t) - v(z,t-h) \\ &= \int_t^T \sum_{z' \in \mathcal{Z}} v(z',s) f(y(s),z')^\top D^{-1}(z',s) p(z',s \mid z,t) \mathrm{d}y(s) \\ &\quad - \int_t^T \sum_{z' \in \mathcal{Z}} v(z',s) f(y(s),z')^\top D^{-1}(z',s) f(y(s),z') p(z',s \mid z,t) \mathrm{d}s \\ &\quad - \int_{t-h}^T \sum_{z' \in \mathcal{Z}} v(z',s) f(y(s),z')^\top D^{-1}(z',s) p(z',s \mid z,t-h) \mathrm{d}y(s) \\ &\quad + \int_{t-h}^T \sum_{z' \in \mathcal{Z}} v(z',s) f(y(s),z')^\top D^{-1}(z',s) f(y(s),z') p(z',s \mid z,t-h) \mathrm{d}s.\end{aligned} \tag{B.3.62}$$

For notational clarity, we first evaluate the difference of the occuring *stochastic* integrals. We define the convenience function $g(z,s) \coloneqq v(z,s) f(y(s),z)^\top D^{-1}(z,s)$. Also, notice that the transition density (of the prior process, with respect to which we compute the expectation) obeys the backward master equation Section 2.1.2.1: with $s > t$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} p(z',s \mid z,t) = - \sum_{z'' \in \mathcal{Z}} \Lambda(z,z'',t) p(z',s \mid z'',t).$$

With this,

$$\begin{aligned}&\int_t^T \sum_{z' \in \mathcal{Z}} g(z',s) p(z',s \mid z,t) \mathrm{d}y(s) \\ &\qquad - \int_{t-h}^T \sum_{z' \in \mathcal{Z}} g(z',s) p(z,s \mid z,t-h) \mathrm{d}y(s) \\ &= \int_t^T \sum_{z' \in \mathcal{Z}} g(z',s) p(z',s \mid z,t) \mathrm{d}y(s) \\ &\qquad - \left(\int_{t-h}^t \sum_{z' \in \mathcal{Z}} g(z',s) p(z',s \mid z,t-h) \mathrm{d}y(s) \right. \\ &\qquad \left. + \int_t^T \sum_{z' \in \mathcal{Z}} g(z',s,t) \left(p(z,s \mid z,t) - \partial_t p(z',s \mid z,t) \mathrm{d}t \right) \mathrm{d}y(s) \right) \\ &= \left(\int_t^T \sum_{z' \in \mathcal{Z}} g(z',s) \partial_t p(z',s \mid z,t) \mathrm{d}y(s) \right) \mathrm{d}t \\ &\qquad - \int_{t-h}^t \sum_{z' \in \mathcal{Z}} g(z',s) p(z',s \mid z,t-h) \mathrm{d}y(s) \\ &= - \left(\int_t^T \sum_{z' \in \mathcal{Z}} g(z',s) \sum_{z'' \in \mathcal{Z}} p(z',s \mid z'',t) \Lambda(z,z'',t) \mathrm{d}y(s) \right) \mathrm{d}t \\ &\qquad - v(z(t),t) f(y(t),z(t))^\top D^{-1}(z(t),t) \mathrm{d}y(t).\end{aligned} \tag{B.3.63}$$

In the last step, two operations were performed at once: first, we applied the definition of the Itô integral *backwards* in time: as detailed in [115], every SDE can be expressed through a backward stochastic process. For

$$\mathrm{d}Y(t) = f(Y(t),t)\mathrm{d}t + Q(Y(t),t)\mathrm{d}W(t),$$

a backward description is given as

$$\mathrm{d}\overleftarrow{Y}(t) = \overleftarrow{f}(Y(t),t)\mathrm{d}t + Q(Y(t),t)\mathrm{d}\overleftarrow{W}(t)$$

with

$$\begin{aligned}
\overleftarrow{f}_i(y,t) &= f_i(y,t) - \frac{1}{p(y,t)}\sum_{j,k}\partial_{y_j}\{p(y,t)Q_{ik}(y,t)Q_{jk}(y,t)\},\\
\mathrm{d}\overleftarrow{W}_i(t) &= \mathrm{d}W_i(t) + \frac{1}{p(y,t)}\sum_{j}\partial_{y_j}\{p(y,t)Q_{jk}(y,t)\}\mathrm{d}t.
\end{aligned}$$

As can be straightforwardly checked, a state-independent dispersion $Q(Y(t),t) = Q(t)$ yields equality between the forward and backward description. Hence, we can readily invert the respective integral. Secondly, we took the limit $h \to 0$, resulting in the usual differential expression.

Next, consider the conventional Lebesgue integrals within Eq. (B.3.62), and again define for convenience
$h(y(s),z,s) \coloneqq v(z,s)f(y(s),z)^\top D^{-1}(z,s)f(y(s),z)$:

$$\begin{aligned}
&\int_t^T \sum_{z'\in\mathcal{Z}} h(y(s),z',s)p(z' \mid z(t))\mathrm{d}s\\
&\qquad - \int_{t-h}^T \sum_{z'\in\mathcal{Z}} h(y(s),z',t-h)p(z',s \mid z,t-h)\mathrm{d}s\\
&= \int_t^T \sum_{z'\in\mathcal{Z}} g(y(s),z',s)p(z',s \mid z,t)\mathrm{d}s\\
&\qquad - \Bigg(\int_{t-h}^t \sum_{z'\in\mathcal{Z}} h(y(s),z',t)p(z',s \mid z,t-h)\mathrm{d}s\\
&\qquad + \int_t^T \sum_{z'\in\mathcal{Z}} h(y(s),z',t)\left(p(z',s \mid z,t) - \partial_t p(z',s \mid z,t)\mathrm{d}t\right)\mathrm{d}s\Bigg)\\
&= \left(\int_t^T \sum_{z_s\in\mathcal{Z}} h(y(s),z',s)\partial_t p(z',s \mid z,t)\mathrm{d}s\right)\mathrm{d}t\\
&\qquad - \int_{t-h}^t \sum_{z'\in\mathcal{Z}} h(y(s),z',t)p(z',s \mid z,t-h)\mathrm{d}s\\
&= -\left(\int_t^T \sum_{z'\in\mathcal{Z}} h(y(s),z',s)\sum_{z''\in\mathcal{Z}} p(z',s \mid z'',t)\Lambda(z',z'',t)\mathrm{d}s\right)\mathrm{d}t\\
&\qquad - v(z,t)f(y(t),z)^\top D^{-1}(z,t)f(y(t),z)\mathrm{d}t. \qquad \text{(B.3.64)}
\end{aligned}$$

Lastly, add the remaining integrals from Eq. (B.3.63) and Eq. (B.3.64):

$$\begin{aligned}
&\sum_{z''\in\mathcal{Z}} \left(- \int_t^T \left[\sum_{z'\in\mathcal{Z}} g(z', s) p(z', s \mid z'', t) \right] \mathrm{d}y(s) \right. \\
&\qquad \left. + \int_t^T \sum_{z'\in\mathcal{Z}} h(y(s), z', s) p(z', s \mid z'', t) \mathrm{d}s \right) \Lambda(z', z'', t)\mathrm{d}t \\
&= \sum_{z''\in\mathcal{Z}} \left[v(z, t) - 1 \right] \Lambda(z', z'', t)\mathrm{d}t \\
&= \sum_{z''\in\mathcal{Z}} v(z, t)\Lambda(z', z'', t)\mathrm{d}t. \qquad \text{(B.3.65)}
\end{aligned}$$

Inserting Eq. (B.3.63), Eq. (B.3.64) and Eq. (B.3.65) into Eq. (B.3.62), we find the differential provided in the main text,

$$\begin{aligned}
\mathrm{d}v(z, t) = -v(z, t) f(y, z, t)^\top D^{-1}(z, t)\mathrm{d}y(t) - \sum_{z'} v(z', t)\Lambda(z, z', t)\mathrm{d}t \\
+ f(y, z, t)^\top D^{-1}(z, t) f(y, z, t) v(z, t)\mathrm{d}t. \quad \text{(B.3.66)}
\end{aligned}$$

B.3.2 *Forward dynamics*

We want to obtain the dynamics of the unnormalized filtering density

$$\tilde{p}_f(z, t) := \mathsf{E}\left[\mathbb{1}(Z(t) = z) G_{[0,t]} \right] \qquad \text{(B.3.67)}$$

and its normalized counterpart $p_f(z, t)$.

In the unnormalized case, we can straightforwardly compute the differential:

$$\begin{aligned}
\mathrm{d}\tilde{p}_f(z, t) &= \mathsf{E}\left[\mathrm{d}\big(\mathbb{1}(Z(t) = z) G_{[0,t]}\big) \right] \qquad \text{(B.3.68)} \\
&= \mathsf{E}\left[\mathrm{d}\mathbb{1}(Z(t) = z) G_{[0,t]} + \mathbb{1}(Z(t) = z)\mathrm{d}G_{[0,t]} + \mathrm{d}\mathbb{1}(Z(t) = z)\mathrm{d}G_{[0,t]} \right].
\end{aligned}$$

The indicator derivative can be readily expressed via the MJP generator,

$$\mathsf{E}[\mathrm{d}\mathbb{1}(Z(t) = z)] = \mathsf{E}\left[\sum_{z'\in\mathcal{Z}} \Lambda(z, z', t) \mathbb{1}(Z(t) = z)\mathrm{d}t \right]. \qquad \text{(B.3.69)}$$

Acknowledging furthermore that $\mathrm{d}t \cdot \mathrm{d}y(t) = 0$, we obtain (with the G-differential evaluated analogously to Eq. (B.3.61)) the Zakai equation:

$$\begin{aligned}
\mathrm{d}\tilde{p}_f(z, t) &= \mathsf{E}\left[\mathrm{d}\mathbb{1}(Z(t) = z) G_{[0,t]} + \mathbb{1}(Z(t) = z)\mathrm{d}G_{[0,t]} \right] \\
&= \sum_{z'\in\mathcal{Z}} \Lambda(z', z, t)\tilde{p}_f(z', t)\mathrm{d}t + \mathsf{E}\left[\mathbb{1}(Z(t) = z) f(y(t), z(t))^\top D^{-1}(z(t), t) \right] \mathrm{d}y(t) \\
&= \sum_{z'\in\mathcal{Z}} \Lambda(z', z, t)\tilde{p}_f(z', t)\mathrm{d}t + \tilde{p}_f(z, t) f(y(t), z(t))^\top D^{-1}(z(t), t)\mathrm{d}y(t). \qquad \text{(B.3.70)}
\end{aligned}$$

To derive the dynamics of the respective normalized quantity

$$p_f(z,t) = \frac{\mathsf{E}\left[\mathbb{1}(Z(t)=z)G(y_{[0,t]}, Z_{[0,t]})\right]}{\mathsf{E}\left[G(y_{[0,t]}, Z_{[0,t]})\right]},$$

consider its denominator. Notice here that

$$\mathrm{d}\ln \mathsf{E}\big[G_{[0,t]}\big] = \frac{\mathrm{d}\mathsf{E}\big[G_{[0,t]}\big]}{\mathsf{E}\big[G_{[0,t]}\big]} - \frac{1}{2}\frac{\operatorname{tr}\left\{\mathrm{d}\mathsf{E}\big[G_{[0,t]}\big]\,\mathrm{d}\mathsf{E}\big[G_{[0,t]}\big]^\top\right\}}{\mathsf{E}\big[G_{[0,t]}\big]^2}. \tag{B.3.71}$$

The quantity $\mathrm{d}\mathsf{E}[G]$ is precisely given by Eq. (B.3.70) upon replacing the indicator by a constant, $\mathbb{1}(Z(t)=z)\to 1$:

$$\mathrm{d}\mathsf{E}\big[G_{[0,t]}\big] = \mathsf{E}\left[G_{[0,t]} f(y(t), z(t))^\top D^{-1}(z(t),t)\right]\mathrm{d}y(t). \tag{B.3.72}$$

As $\mathrm{d}y(t)\mathrm{d}y(t)^\top = D(z(t),t)\mathrm{d}t$, we find upon inserting this into Eq. (B.3.71)

$$\begin{aligned}\mathrm{d}\ln \mathsf{E}\big[G_{[0,t]}\big] &= \frac{\mathsf{E}\big[G_{[0,t]} f(y(t),z(t))^\top D^{-1}(z(t),t)\big]}{\mathsf{E}\big[G_{[0,t]}\big]}\mathrm{d}y(t)\\ &- \frac{1}{2}\frac{\mathsf{E}\big[G_{[0,t]} f(y(t),z(t))^\top D^{-1}(z(t),t)\big]\, D^{-1}(z(t),t)\,\mathsf{E}[D^{-1}(z(t),t) f(y(t),z(t))]}{\mathsf{E}\big[G_{[0,t]}\big]^2}\mathrm{d}t.\end{aligned} \tag{B.3.73}$$

Notice that the terms on the right-hand side are of the same form as the Radon-Nikodym derivative G, cf. Eq. (B.3.61). We can write

$$\mathsf{E}[G] = \exp\left\{\int_0^t \underbrace{\frac{\mathsf{E}\big[Gf^\top D^{-1}\big]}{\mathsf{E}[G]}}_{=:\varpi^\top}\mathrm{d}y(s) - \frac{1}{2}\int_0^t \underbrace{\frac{\mathsf{E}\big[Gf^\top D^{-1}\big]\,D\,\mathsf{E}[D^{-1}fG]}{\mathsf{E}[G]^2}}_{\varpi^\top D^{-1}\varpi}\mathrm{d}s\right\}, \tag{B.3.74}$$

where, for clarity, we restate with arguments:

$$\varpi(t) = \frac{\mathsf{E}\big[G(y_{[0,t]}, Z_{[0,t]}) f(y(t), Z(t))^\top D(Z(t),t)^{-1}\big]}{\mathsf{E}\big[G(y_{[0,t]}, Z_{[0,t]})\big]}. \tag{B.3.75}$$

Inserting this into the above normalizer expression for $p_f(z,t)$, we arrive at

$$\begin{aligned}p_f(z,t) = \mathsf{E}\Bigg[\mathbb{1}(Z(t)=z)\exp\Bigg\{&\int_0^t \left(f^\top(y(s),Z(s))D(Z(s))^{-1} - \varpi\right)(\mathrm{d}y(s) - \varpi(s))\\ -\frac{1}{2}\int_0^t &\left(f^\top(y(s),Z(s)) - \varpi(s)\right) D(Z(s))^{-1}\left(f(y(s),Z(s)) - \varpi(s)\right)\mathrm{d}s\Bigg\}\Bigg].\end{aligned} \tag{B.3.76}$$

In the same way as in the above derivation of the Zakai equation, one can compute the differentials for this expression, yielding the *Kushner-Stratonovich* equation

$$\begin{aligned}\mathrm{d}p_f(z,t) &= \sum_{z'\in\mathcal{Z}}\Lambda(z',z,t)p_f(z,t)\mathrm{d}t\\ &\quad + p_f(z,t)\left(f(y(t), z - \bar{f}(y,t))\right)^\top D(z,t)^{-1}\left(\mathrm{d}y(t) - \bar{f}(y,t)\right),\end{aligned} \tag{B.3.77}$$

where $\bar{f}(y,t) := \sum_{z\in\mathcal{Z}} f(y, z(t))p_f(z,t)$.

B.3.3 *Smoothing dynamics*

To obtain the dynamics of the unnormalized smoothing distribution

$$\mathrm{d}\tilde{p}_s(z,t) = v(z,t)\mathrm{d}\tilde{p}_f(z,t) + \tilde{p}_f(z,t)\mathrm{d}v(z,t) + \mathrm{d}\tilde{p}_f(z,t)\mathrm{d}v(z,t),$$

the only missing ingredient is the cross term, for which we find

$$\mathrm{d}\tilde{p}_f(z,t)\mathrm{d}v(z,t) = -\tilde{p}_f(z,t)v(z,t)f(y(t),z)^\top D^{-1}(t)f(y(t),z)\mathrm{d}t.$$

With these individual results, we finally obtain the dynamics as

$$\begin{aligned}
\tilde{p}_s(z,t) &= v(z,t)\mathrm{d}\tilde{p}_f(z,t) + \tilde{p}_f(z,t)\mathrm{d}v(z,t) + \mathrm{d}\tilde{p}_f(z,t)\mathrm{d}v(z,t) \\
&= v(z,t)\left(\sum_{z'\in\mathcal{Z}} \Lambda(z',z,t)\tilde{p}_f(z',t)\mathrm{d}t + \tilde{p}_f(z,t)f(y(t),z)^\top D^{-1}(z,t)\mathrm{d}y(t)\right) \\
&\quad + \tilde{p}_f(z,t)\Bigg(-v(z,t)f(y,z)^\top D^{-1}(z,t)\mathrm{d}y(t) - \sum_{z'} v(z',t)\Lambda(z,z',t)\mathrm{d}t \\
&\qquad + v(z,t)f(y,z)^\top D^{-1}(t)f(y,z)\mathrm{d}t\Bigg) \\
&\quad - \tilde{p}_f(z,t)v(z,t)f(y(t),z)^\top D^{-1}(z,t)f(y(t),z)\mathrm{d}t \\
&= v(z,t)\sum_{z'} \Lambda(z',z,t)\tilde{p}_f(z',t)\mathrm{d}t - \sum_{z'} v(z',t)\Lambda(z,z',t)\tilde{p}_f(z,t)\mathrm{d}t.
\end{aligned}$$

B.4 EXPERIMENTAL DETAILS

B.4.1 *Synthetic data generation*

The 1D synthetic data are generated with the following parameters; note that we provide z-dependent parameters as vectors in the 1D case, where, e.g., $A(z) = A_z$.

$$\begin{aligned}
\pi_{z_0} &= (1,0)^\top & A &= (-0.5,-0.5)^\top \\
\mu_0 &= 0 & b &= (1,-1)^\top \\
\Sigma_0 &= 0.1 & D &= (0.1,0.1)^\top \\
\Lambda &= \begin{pmatrix} -0.2 & 0.2 \\ 0.2 & -0.2 \end{pmatrix}, & \Sigma_x &= 0.05
\end{aligned} \tag{B.4.78}$$

The 2D synthetic data are generated from

$$\begin{aligned}
\pi_{z_0} &= (1,0)^\top & A &= \begin{pmatrix} -0.1 & 1.4 \\ -2.6 & 0.6 \end{pmatrix}, \begin{pmatrix} 0.6 & -1.4 \\ 2.6 & 0.6 \end{pmatrix} \\
\mu_0 &= (1,0)^\top, (-1,0)^\top & b &= (1,0)^\top, (-1,0)^\top \\
\Sigma_0 &= \begin{pmatrix} 0.49 & 0 \\ 0 & 0.49 \end{pmatrix} & D &= \begin{pmatrix} 0.49 & 0 \\ 0 & 0.49 \end{pmatrix} \\
\Lambda &= \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}, & \Sigma_x &= \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}
\end{aligned} \tag{B.4.79}$$

The same observation model parameterization was used for both modes z.

B.4.2 *Hyperparameter settings*

When performing inference, we initialize the distribution hyperparameters empirically. To this end, we run k-means with the number of modes $|\mathcal{Z}|$ on the data and obtain empirical cluster means μ_z and covariances Σ_z. While we make use of μ_z, we do not utilize Σ_z, however, as it does not contain information about the temporal evolution about the process. Instead, we re-compute Σ_z as an empirical estimate of the quadratic variation of the process:

$$\Sigma_z = \frac{\sum_{i=1}^N \mathbb{1}(z(t_i) = z) \frac{\Delta x_i^\top \Delta x_i}{\Delta t_i}}{\sum_{i=1}^N \mathbb{1}(z(t_i) = z)}.$$

In the following, we denote by $z(t_i)$ the k-means cluster assignment of observation i at time point t_i.

INITIAL CONDITIONS The MJP initial Dirichlet hyperparameters

$$\alpha_z = 1 + \delta_{z(t_1)}.$$

The SDE initial NIW hyperparameters

$$\eta = \frac{\sum_z \mu_z}{|Z|}, \quad \lambda = 1, \quad \Psi = 0.1\frac{\sum_z \Sigma_z}{|Z|}, \quad \kappa = n + 2.$$

Note that - as done, e.g., in [96] - we use a heuristic downscaling of the empirical covariances as they contain contributions by the measurement noise, the process covariance as well as the drift. Also, $\kappa = n + 2$ is the smallest scaling parameter that makes the IW distribution well defined.

MJP RATES In the 1D case, we compute the number of total observed transitions in the k-means trajectory, N_{trans} and set the Gamma hyperparameters as

$$s = N_{\text{trans}}, \quad r = 1.$$

As we have non-stationary dynamics in the 2D example, we here simply set

$$s = r = 1.$$

SDE DRIFT PARAMETERS In the 1D example, we compute

$$\begin{aligned} \hat{A}_z &= \sum_{i=1}^N \mathbb{1}(z(t_i) = z)\frac{x_{i+1} - x_i}{x_{i+1} - t_i}, \\ \hat{b}_z &= -\hat{A}_z \mu_z, \end{aligned} \tag{B.4.80}$$

where the latter is because the set point for a linear system is found via

$$f(y) = Ay + b = A(y + A^{-1}b).$$

Hence $f(y) = 0$ if $y = -A^{-1}b$, and we demand

$$\mu_z = -A_z^{-1}b \Rightarrow b_z = -A_z\mu_z.$$

With this, the MN hyperparameters

$$M_z = [\hat{A}_z, \hat{b}_z], \quad K_z = \mathbb{1}_{n+1}.$$

In the 2D example, we follow the same procedure, but set manually

$$\hat{A}_z = -\,\mathbb{1}_2, \quad K_z = 0.001\,\mathbb{1}_3,$$

with the 2- and 3-dimensional identity matrices.

SDE DISPERSION In the 1D example, we set $D(z) = 0.1\Sigma_z$, in the 2D example (due to the much larger variability) $D(z) = 0.05\Sigma_z$ with a heuristic downscaling as above.

OBSERVATION COVARIANCE Lastly,

$$\Psi_x = 0.5\Sigma_z \quad \lambda_{D_z} = n + 2$$

in the 1D and

$$\Psi_x = 0.0025\Sigma_z \quad \lambda_{D_z} = n + 2$$

in the 2D case.

C

APPENDIX C: VARIATIONAL INFERENCE FOR HYBRID SYSTEMS

C.1 ALTERNATIVE DERIVATION OF THE HYBRID PROCESS KL DIVERGENCE

To obtain $\mathsf{D}_{\mathrm{KL}}[\mathsf{Q} \mid\mid \mathsf{P}]$ for two MJP-SSDE processes, consider discretized versions of the continuous processes on a regular time grid, $t \in \{0, h, 2h, \dots, K \cdot h = T\}$ for some small h, where we aim to take the limit $h \to 0$ of the resulting expressions. Abbreviating $Y_k = Y(k \cdot h)$ and Z_k analogously, we have the discretized paths $\{Y_k, Z_k\}_{k \in \{0,1,\dots,K\}}$. For these joint paths, one can explicitly write down the probability density functions: we have

$$q(y_{[0,K]}, z_{[0,K]}) = q(y_0, z_0) \prod_{k=1}^{K} q(y_k, z_k \mid y_{k-1}, z_{k-1}),$$
$$p(y_{[0,K]}, z_{[0,K]}) = p(y_0, z_0) \prod_{k=1}^{K} p(y_k, z_k \mid y_{k-1}, z_{k-1}).$$

Insertion into the KL divergence yields

$$\mathsf{D}_{\mathrm{KL}}[q \mid\mid p] = \mathsf{D}_{\mathrm{KL}}[q^0 \mid\mid p^0]$$
$$+\sum_{k=1}^{K} \sum_{z_0,\dots,z_K} \int \left(q(y_0, z_0) \prod_{i=1}^{k} q(y_i, z_i \mid y_{i-1}, z_{i-1}) \ln \frac{q(y_i, z_i \mid y_{i-1}, z_{i-1})}{p(y_i, z_i \mid y_{i-1}, z_{i-1})} \right) \mathrm{d}y_0 \cdots \mathrm{d}y_k, \tag{C.1.1}$$

where

$$\mathsf{D}_{\mathrm{KL}}[q^0 \mid\mid p^0] = \sum_{z_0} \int q(y_0, z_0) \ln \frac{q(y_0, z_0)}{p(y_0, z_0)} \mathrm{d}y_0 \tag{C.1.2}$$

denotes the KL divergence of the initial distributions.

Any of the K summands from the second term can be simplified as

$$
\begin{aligned}
&\sum_{z_0,\dots,z_K} \int \left(q(y_0, z_0) \prod_{i=1}^{k} q(y_k, z_k \mid y_{k-1}, z_{k-1}) \right) \ln \frac{q(y_k, z_k \mid y_{k-1}, z_{k-1})}{p(y_k, z_k \mid y_{k-1}, z_{k-1})} \mathrm{d}y_0 \cdots \mathrm{d}y_K \\
&= \sum_{z_{k-1}, z_k} \int q(y_k, z_k \mid y_{k-1}, z_{k-1}) q(y_{k-1}, z_{k-1}) \ln \frac{q(y_k, z_k \mid y_{k-1}, z_{k-1})}{p(y_k, z_k \mid y_{k-1}, z_{k-1})} \mathrm{d}y_{k-1} \mathrm{d}y_k \\
&= \sum_{z_{k-1}, z_k} \int q(y_k \mid z_k, y_{k-1}, z_{k-1}) q(z_k \mid y_{k-1}, z_{k-1}) q(y_{k-1}, z_{k-1}) \\
&\qquad \left(\ln \frac{q(y_k \mid z_k, y_{k-1}, z_{k-1})}{p(y_k \mid z_k, y_{k-1}, z_{k-1})} + \ln \frac{q(z_k \mid y_{k-1}, z_{k-1})}{p(z_k \mid y_{k-1}, z_{k-1})} \right) \mathrm{d}y_{k-1} \mathrm{d}y_k . \qquad \text{(C.1.3)}
\end{aligned}
$$

The model structure further implies

$$q(z_k \mid y_{k-1}, z_{k-1}) = q(z_k \mid z_{k-1}),$$

as there is no feedback from the Y- to the Z-process. All summands hence decompose into an SSDE and an MJP contribution.

SSDE-KL CONTRIBUTION Inspecting the first part of Eq. (C.1.3) and utilizing again the expansion (cf. Appendix B.1.2)

$$q(z_k \mid z_{k-1}) = \delta_{z_k, z_{k-1}} + \tilde{\Lambda}_{z_{k-1}, z_k} h + o(h),$$

we have

$$
\begin{aligned}
&\sum_{z_{k-1}, z_k} \int q(y_k \mid y_{k-1}, z_{k-1}) \left(\delta_{z_k, z_{k-1}} + \tilde{\Lambda}_{z_{k-1}, z_k} h + o(h) \right) \\
&\qquad \cdot q(y_{k-1}, z_{k-1}) \left(\ln \frac{q(y_k \mid y_{k-1}, z_{k-1})}{p(y_k \mid y_{k-1}, z_{k-1})} \right) \mathrm{d}y_{k-1} \mathrm{d}y_k \\
&= \sum_{z_{k-1}, z_k} \int q(y_k \mid y_{k-1}, z_{k-1}) \delta_{z_k, z_{k-1}} q(y_{k-1}, z_{k-1}) \left(\ln \frac{q(y_k \mid y_{k-1}, z_{k-1})}{p(y_k \mid y_{k-1}, z_{k-1})} \right) \mathrm{d}y_{k-1} \mathrm{d}y_k \\
&+ \sum_{z_{k-1}, z_k} \int q(y_k \mid y_{k-1}, z_{k-1}) \tilde{\Lambda}_{z_{k-1}, z_k} h \left(\ln \frac{q(y_k \mid y_{k-1}, z_{k-1})}{p(y_k \mid y_{k-1}, z_{k-1})} \right) \mathrm{d}y_{k-1} \mathrm{d}y_k + o(h).
\end{aligned}
$$

Employing furthermore the Euler-Maruyama approximation

$$q(y_k \mid z_k, y_{k-1}, z_{k-1}) = \mathcal{N}(y_k \mid y_{k-1} + g(y_{k-1}, z_{k-1}, (k-1)h)h, D(z_{k-1})h)$$

and analogously for $p(y_k \mid z_k, y_{k-1}, z_{k-1})$, we find for the log-fraction

$$
\begin{aligned}
\ln \frac{q(y_k \mid y_{k-1}, z_{k-1})}{p(y_k \mid y_{k-1}, z_{k-1})} &= \frac{-\frac{1}{2h} \| y_k - y_{k-1} - g(y_{k-1}, z_{k-1}, (k-1)h) \cdot h \|_{D^{-1}}^2}{-\frac{1}{2h} \| y_k - y_{k-1} - f(y_{k-1}, z_{k-1}, (k-1)h) \cdot h \|_{D^{-1}}^2} \\
&= \frac{h}{2} \cdot \| g(y_{k-1}, z_{k-1}, (k-1)h) - f(y_{k-1}, z_{k-1}, (k-1)h) \|_{D^{-1}}^2
\end{aligned}
$$

with the shorthand $\|x\|_A^2 := x^\top A x$. This ratio does not depend on the time step k, but only $k-1$; consequently, z_k, y_k can be marginalized out.

Note that the Gaussian normalizing prefactor $(2\pi)^{\frac{k}{2}}|D|^{-\frac{1}{2}}$ is the same for both distributions and hence cancels; if Q and P had different dispersions D_{Q} and D_{P}, this cancellation would not occur and the KL would diverge [106]. This is a reflection of the fact that the laws of two SDEs with different dispersions are singular with respect to each other. Taking the limit $K \to \infty$, $h \to 0$ with $K \cdot h = T$ yields an integral expression:

$$\begin{aligned}\int_0^T \sum_z \int q(y,z,t)\frac{1}{2}\|g(y,z,t) - f(y,z,t)\|^2_{D^{-1}}\,\mathrm{d}y\mathrm{d}t& \\ = \frac{1}{2}\int_0^T \mathsf{E}\left[\|g(Y(t),Z(t),t) - f(Y(t),Z(t),t)\|^2_{D^{-1}}\right]\mathrm{d}t.&\end{aligned}$$

MJP-KL CONTRIBUTION As the second term of Eq. (C.1.3) does not depend on the Y-process, it is simply the KL divergence between two MJPs. Its derivation is analogous to the one presented above and can be found, e.g., in [42].

C.2 OPTIMIZING THE VARIATIONAL PARAMETERS

We provide here the explicit gradients with respect to all variational parameters. We have

$$\begin{aligned}\partial_{A_q(z,t)}\mathscr{L} &= -\frac{1}{2}\partial_{A_q(z,t)}\mathsf{E}\left[\|g-f\|^2_{D^{-1}}\right] - \lambda^\top(z,t)\mu(z,t) - 2\Psi(z,t)\Sigma(z,t) \\ &= -q_Z(z,t)D^{-1}\left(\bar{A}(z,t)(\mu(z,t)\mu^\top(z,t) + \Sigma(z,t)) + \bar{b}(z,t)\mu^\top(z,t)\right) \\ &\quad - \lambda^\top(z,t)\mu(z,t) - 2\Psi(z,t)\Sigma(z,t).\end{aligned} \tag{C.2.4}$$

Similarly, we find

$$\begin{aligned}\partial_{b_q(z,t)}\mathscr{L} &= -\partial_{b_q(z,t)}\frac{1}{2}\mathsf{E}\left[\|g-f\|^2_{D^{-1}}\right] - \lambda(z,t) \\ &= -q_Z(z,t)D^{-1}\left(\bar{A}(z,t)\mu(z,t) + \bar{b}(z,t)\right) - \lambda(z,t).\end{aligned} \tag{C.2.5}$$

Finally,

$$\begin{aligned}\partial_{\tilde{\Lambda}_{zz'}(t)}\mathscr{L} &= -\partial_{\tilde{\Lambda}_{zz'}(t)}\left[\sum_{z'\in\mathcal{Z}\setminus z}\left\{\tilde{\Lambda}(z,z',t)\ln\frac{\tilde{\Lambda}(z,z',t)}{\Lambda(z,z',t)}\right\} - (\tilde{\Lambda}(z,t) - \Lambda(z,t))\right] \\ &\quad + \nu(z,t)q_Z(z,t) - \nu(z',t)q_Z(z,t) \\ &= q_Z(z,t)\left(-\ln\frac{\tilde{\Lambda}_{zz'}(t)}{\Lambda_{zz'}} + \nu(z,t) - \nu(z',t)\right).\end{aligned}$$

The gradients with respect to the initial conditions result from Pontryagin's maximum principle [126, 127] as

$$
\begin{aligned}
\partial_{\mu(z,0)}\mathscr{L} &= \partial_{\mu(z,0)}\,\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0 \parallel \mathsf{P}^0\big] + \lambda(z,0) = 0,\\
\partial_{\Sigma(z,0)}\mathscr{L} &= \partial_{\Sigma(z,0)}\,\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0 \parallel \mathsf{P}^0\big] + \Psi(z,0) = 0,\\
\partial_{q_Z(z,0)}\mathscr{L} &= \partial_{q_Z(z,0)}\,\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0 \parallel \mathsf{P}^0\big] + \nu(z,0) = 0.
\end{aligned}
\tag{C.2.6}
$$

In principle, one could use these expressions to find closed-form solutions for the initial parameters. A direct reset of the parameters may however cause numerical instabilities in the forward-backward sweeping algorithm. We hence utilize the same gradient ascent update scheme as above.

We assume a Gaussian prior initial distribution, i.e. $p(y,0 \mid z) = \mathcal{N}(y \mid \mu_p^0(z), \Sigma_p^0(z))$, and for the presented variational ansatz we have an initial variational distribution $q(y,0 \mid z) = \mathcal{N}(y \mid \mu(z,0), \Sigma(z,0))$, which is also Gaussian. This yields

$$
\begin{aligned}
\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z} \parallel \mathsf{P}^0_{Y|Z}\big] &= \mathsf{D}_{\mathrm{KL}}\big[\mathcal{N}(y \mid \mu(z,0), \Sigma(z,0)) \parallel \mathcal{N}(y \mid \mu_p^0(z), \Sigma_p^0(z))\big]\\
&= \frac{1}{2}\Big\{\ln \frac{|\Sigma_p^0(z)|}{|\Sigma(z,0)|} + \operatorname{tr}\big(\Sigma_p^0(z)^{-1}\Sigma(z,0)\big)\\
&\quad + \big(\mu_p^0(z) - \mu(z,0)\big)\,\Sigma_p^0(z)^{-1}\,\big(\mu_p^0(z) - \mu(z,0)\big)^\top + n\Big\}.
\end{aligned}
$$

We readily compute

$$
\begin{aligned}
\partial_{\mu(z,0)}\,\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y,Z} \parallel \mathsf{P}^0_{Y,Z}\big] &= q_Z(z,0)\Sigma_p^0(z)^{-1}(\mu(z,0) - \mu_p^0(z)),\\
\partial_{q_Z(z,0)}\,\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y,Z} \parallel \mathsf{P}^0_{Y,Z}\big] &= \partial_{q_Z(z,0)}\,\big\{\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Z} \parallel \mathsf{P}^0_{Z}\big]\big\} + \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z} \parallel \mathsf{P}^0_{Y|Z}\big].
\end{aligned}
\tag{C.2.7}
$$

For the covariance matrix $\Sigma(z,0)$ and the initial distribution $q_Z(z,0)$, we require additional constraints. The covariance $\Sigma(z,0)$ needs to be positive semi-definite, which can be enforced by a reparameterization as $\Sigma(z,0) = CC^\top$. We compute the gradient of

$$
\begin{aligned}
\mathscr{L}(C) = q(z,0)\,\big\{\mathsf{D}_{\mathrm{KL}}\big[\mathcal{N}(y \mid \mu(z,0), CC^\top) \parallel \mathcal{N}(y \mid \mu_p^0(z), \Sigma_p^0(z))\big]&\\
+ \operatorname{tr}\big(\Psi(z,0)^\top CC^\top\big)\big\}&
\end{aligned}
$$

utilizing the PyTorch package for automatic differentiation and optimization [220].

The initial distribution $q_Z(z,t)$ needs to fulfil $\sum_z q_Z(z,0) = 1$, so we optimize an augmented cost function

$$
\begin{aligned}
\mathscr{L}(\mathsf{Q}^0_Z, \xi) = \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_Z \parallel \mathsf{P}^0_Z\big] + \sum_z q_Z(z,0)\,\mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z} \| \mathsf{P}^0_{Y|Z}\big]&\\
+ \sum_z \nu(z,0) q_Z(z,0) + \xi(1 - \sum_z q_Z(z,0)),&
\end{aligned}
$$

where $\xi(z)$ are Lagrange multipliers. We can again eliminate the constraints by enforcing a reparameterization [221] as $q_Z(z,0) = q_z$ for $z \in \{1, \dots, k-1\}$, with $k = |\mathcal{Z}|$ and $q_Z(k,0) = 1 - \sum_{z=1}^{k-1} q_z$, yielding the unconstrained problem

$$\begin{aligned}
\mathscr{L}(q_1, \dots, q_{k-1}) = &\sum_{z=1}^{k-1} q_z \ln \frac{q_z}{p(z,0)} + (1 - \sum_{z=1}^{k-1} q_z) \ln \frac{1 - \sum_{z=1}^{k-1} q_z}{p(k,0)} \\
&+ \sum_{z=1}^{k-1} q_z \, \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z=z} \,\big\|\, \mathsf{P}^0_{Y|Z=z}\big] \\
&+ (1 - \sum_{z=1}^{k-1} q_z) \, \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z=k} \,\big\|\, \mathsf{P}^0_{Y|Z=k}\big] \\
&+ \sum_{z=1}^{k-1} \nu(z,0) q_z + \nu(k,0)(1 - \sum_{z=1}^{k-1} q_z).
\end{aligned}$$

We find

$$\begin{aligned}
\partial_{q_z} \mathscr{L} &= \ln \frac{q_z}{p(z,0)} + 1 - \ln \frac{1 - \sum_{z=1}^{k-1} q_z}{p(k,0)} - 1 \\
&\quad + \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z=z} \,\big\|\, \mathsf{P}^0_{Y|Z=z}\big] - \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z=k} \,\big\|\, \mathsf{P}^0_{Y|Z=k}\big] + \nu(z,0) - \nu(k,0) \\
&= \ln \frac{q_z p(k,0)}{(1 - \sum_{z=1}^{k-1} q_z) p(z,0)} \qquad \text{(C.2.8)} \\
&\quad + \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z=z} \,\big\|\, \mathsf{P}^0_{Y|Z=z}\big] - \mathsf{D}_{\mathrm{KL}}\big[\mathsf{Q}^0_{Y|Z=k} \,\big\|\, \mathsf{P}^0_{Y|Z=k}\big] + \nu(z,0) - \nu(k,0).
\end{aligned}$$

C.3 OPTIMIZING THE PRIOR PARAMETERS

The parameters of the original process P can as well be learned straightforwardly by optimizing the full Lagrangian via gradient ascent.

PRIOR MJP TRANSITION RATES With the usual shorthand $\Lambda_{zz'}(t) = \Lambda(z, z', t)$, we compute the prior transition rates (which we in all cases assume to be time-homogeneous, $\Lambda_{zz'}(t) = \Lambda_{zz'}$):

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \Lambda_{ij}} &= -\frac{\partial}{\partial \Lambda_{ij}} \int_0^T \sum_z q_Z(z,t) \sum_{z' \in \mathcal{Z} \setminus z} \left\{ \tilde{\Lambda}_{zz'}(t) \ln \frac{\tilde{\Lambda}_{zz'}(t)}{\Lambda_{zz'}} \right\} - (\tilde{\Lambda}(z,t) - \Lambda(z)) \mathrm{d}t \\
&= \frac{\partial}{\partial \Lambda_{ij}} \int_0^T \sum_z q_Z(z,t) \left[\sum_{z' \in \mathcal{Z} \setminus z} \left\{ \tilde{\Lambda}_{zz'}(t) \ln \Lambda_{zz'} - \Lambda_{zz''} \right\} \right] \mathrm{d}t \\
&= \frac{1}{\Lambda_{ij}} \int_0^T q_Z(i,t) \tilde{\Lambda}_{ij}(t) \mathrm{d}t - \int_0^T q_Z(i,t) \mathrm{d}t. \qquad \text{(C.3.9)}
\end{aligned}$$

Setting this to zero yields

$$\Lambda_{ij} = \frac{\int_0^T q_Z(i,t)\tilde{\Lambda}_{ij}(t)\mathrm{d}t}{\int_0^T q_Z(i,t)\mathrm{d}t}. \tag{C.3.10}$$

OBSERVATION COVARIANCE To determine the observation covariance, we compute

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \Sigma_x^{-1}} &= -\frac{\partial}{\partial \Sigma_x^{-1}} \mathsf{E}\left[\sum_i \ln p(x_i \mid y_i)\right] \qquad \text{(C.3.11)}\\ &= -\frac{N}{2}\Sigma_x + \sum_{i=1}^{N} \frac{1}{2} \sum_{z\in\mathcal{Z}} q_Z(z,t_i) \left[(x_i - \mu(z,t_i))(x_i - \mu(z,t_i))^\top + \Sigma(z,t_i)\right],\end{aligned}$$

yielding

$$\Sigma_x = \frac{1}{N}\sum_{i=1}^{N}\sum_{z\in\mathcal{Z}} q_Z(z,t_i)\left[(x_i - \mu(z,t_i))(x_i - \mu(z,t_i))^\top + \Sigma(z,t_i)\right]. \tag{C.3.12}$$

DISPERSION The gradient with respect to the dispersion

$$\begin{aligned}\partial_D \mathscr{L} &= \partial_D \frac{1}{2}\int_0^T \mathsf{E}\left[\|g - f\|^2_{D^{-1}}\right]\mathrm{d}t + \partial_D \sum_z \int_0^T \operatorname{tr}\{\Psi^\top(z,t) D\}\mathrm{d}t\\ &= \frac{1}{2}\int_0^T \partial_D \mathsf{E}\left[\|g - f\|^2_{D^{-1}}\right]\mathrm{d}t + \int_0^T \sum_{z\in\mathcal{Z}} \Psi(z,t)\mathrm{d}t\\ &= \frac{1}{2} - D^{-\top}\left(\int_0^T \sum_{z\in\mathcal{Z}} q_Z(z,t)\, \mathsf{E}[(\bar{A}(z,t)y + \bar{b})(\bar{A}(z,t)y + \bar{b})^\top | z]\right) D^{-\top} \qquad \text{(C.3.13)}\\ &\qquad + \int_0^T \sum_{z\in\mathcal{Z}} \Psi(z,t)\mathrm{d}t.\end{aligned}$$

Note that the more general mode-dependent dispersion $D(z)$ are found in the same way by omitting the summation over z. The prior initial conditions $\mu_p^0(z), \Sigma_p^0(z), p(z,0)$ trivially minimize their KL divergence to the variational initial conditions by equality.

PRIOR DRIFT PARAMETERS The gradients with respect to the model slope and intercept $A(z)$ and $b(z)$ are found as

$$\begin{aligned}\partial_{A(z)}\mathscr{L} &= \partial_{A(z)}\frac{1}{2}\int_0^T \mathsf{E}\left[\|g-f\|^2_{D^{-1}}\right]\mathrm{d}t \qquad (\text{C.3.14})\\ &= \frac{1}{2}\int_0^T q(z,t)\partial_{A(z)}\left\{\operatorname{tr}\{\bar{A}(z,t)^\top D^{-1}\bar{A}(z,t)\Sigma(z,t)\}\right.\\ &= -\int_0^T q_Z(z,t)\left(D^{-1}\bar{A}(z,t)\left(\Sigma(z,t)+\mu(z,t)\mu^\top(z,t)\right)+\right.\\ &\qquad\qquad \left.D^{-1}\bar{b}(z,t)\mu^\top(z,t)\right)\mathrm{d}t,\\ \partial_{b(z)}L &= \partial_{b(z)}\frac{1}{2}\int_0^T \mathsf{E}\left[\|g-f\|^2_{D^{-1}}\right]\mathrm{d}t \qquad (\text{C.3.15})\\ &= \int_0^T q_Z(z,t)D^{-1}(\bar{A}(z,t)\mu(z,t)+\bar{b}(z,t))\end{aligned}$$

C.4 EXPERIMENTAL DETAILS

C.4.1 *Synthetic data generation*

Figure 4.3 is generated with the same parameter settings as the 1D example in Chapter 3. Figure 4.2 is the same except for the slope A, which here is $A(z) = (-1,-1)^\top$.

Figure 4.4 is generated using slightly different parameters:

$$\begin{aligned} \pi_{z_0} &= (1,0)^\top & A &= (-1.5,-1.5)^\top \\ \mu_0 &= (1,-1)^\top & b &= (1.5,-1.5)^\top \\ \Sigma_0 &= (0.2,0.2)^\top & D &= (0.25,0.25)^\top \qquad (\text{C.4.16})\\ \Lambda &= \begin{pmatrix}-0.2 & 0.2\\ 0.2 & -0.2\end{pmatrix}, & \Sigma_x &= 0.1\end{aligned}$$

In the multi-well diffusion examples, we set $\Sigma_{\text{obs}} = 0.0225$ (in 1D), and $\Sigma_{\text{obs}} = \begin{pmatrix}0.2 & 0\\ 0 & 0.2\end{pmatrix}$ in 2D.

C.4.2 *Hyperparameter initialization*

We again utilize k-means to initialize the inference algorithm to obtain empirical cluster means μ_z. As before, we compute

$$\Sigma_z = \frac{\sum_{i=1}^N \mathbb{1}(z(t_i)=z)\frac{\Delta x_i^\top \Delta x_i}{\Delta t_i}}{\sum_{i=1}^N \mathbb{1}(z(t_i)=z)}$$

as an empirical estimate of the quadratic variation. As the VI algorithm in principle depends on the initial conditions, we initialize the variational distributions similar to the hyperparameters of the parameter distributions used in Chapter 3.

Parameter	Learned value
$A(z)$	$(1.11, 0.94, 0.99, 0.99)^\top$
$b(z)$	$(0.88, -0.25, 0.24, -0.68)^\top$
Λ	$\begin{pmatrix} -0.69 & 0.14 & 0.48 & 0.07 \\ 0.07 & -1.39 & 0.45 & 0.87 \\ 1.11 & 0.87 & -2.33 & 0.35 \\ 0.04 & 1.44 & 0.50 & -1.98 \end{pmatrix}$
D	$[0.015, 0.001, 0.003, 0.01]$
$\mu_p(z, 0)$	$[0.99, -0.24, 0.25, -0.69]$
$\Sigma_p(z, 0)$	$[0.002, 0.009, 0.014, 0.025]$
$p(z, 0)$	$[0.92, 0, 0.007, 0.073]$

TABLE C.1: Parameters learned from 1D, four-well SDE data

INITIAL CONDITIONS The initial variational MJP marginals are set as

$$q_Z(z, 0) = 0 + \delta_{z(t_0)}.$$

The GP counterparts

$$\mu(z, 0) = \mu_z, \quad \Sigma(z, 0) = 0.2\Sigma_z,$$

which, for the first iteration, also are set to be the mean and covariance function of the full GPs. We employ heuristic downscaling as before, cf. Appendix B.4 and [96].

MJP RATES We set

$$\tilde{\Lambda}(z, z') = |\mathcal{Z}|^{-1} \quad \forall z' \neq z.$$

SDE DRIFT PARAMETERS In 1D, we proceed as before in Appendix B.4. In 2D, we initialize manually

$$A_q(z) = -\mathbb{1}$$

to ensure numerical stability, as starting with non-contractive dynamics would cause the system to diverge right at the start.

SDE DISPERSION The dispersion $D(z)$ is set to the GP covariance $\Sigma(z, 0)$.

OBSERVATION COVARIANCE Lastly, $\Sigma_x = 0.5\Sigma(z, 0)$.

Parameter	Learned value
A_z	$\begin{pmatrix} -0.91 & 0.03 \\ -0.05 & -1.02 \end{pmatrix}, \begin{pmatrix} -1.07 & -0.04 \\ -0.09 & -0.99 \end{pmatrix}, \begin{pmatrix} -1.01 & 0.04 \\ -0.02 & -1.05 \end{pmatrix}$
b_z	$\begin{pmatrix} 0.98 \\ 0.34 \end{pmatrix}, \begin{pmatrix} -0.83 \\ 0.10 \end{pmatrix}, \begin{pmatrix} -0.02 \\ 1.16 \end{pmatrix}$
Λ	$\begin{pmatrix} -0.75 & 0.29 & 0.46 \\ 0.26 & -0.75 & 0.49 \\ 0.80 & 0.77 & -1.57 \end{pmatrix}$
$D(z)$	$\begin{pmatrix} 0.048 & -0.003 \\ -0.003 & 0.077 \end{pmatrix}, \begin{pmatrix} 0.088 & -0.035 \\ -0.035 & 0.092 \end{pmatrix}, \begin{pmatrix} 0.077 & -0.004 \\ -0.004 & 0.040 \end{pmatrix}$
$\mu_p(z,0)$	$\begin{pmatrix} 1.18 \\ 0.11 \end{pmatrix}, \begin{pmatrix} -0.41 \\ -0.01 \end{pmatrix}, \begin{pmatrix} -0.54 \\ 0.03 \end{pmatrix}$
$\Sigma_p(z,0)$	$\begin{pmatrix} 0.263 & 0.058 \\ 0.058 & 0.269 \end{pmatrix}, \begin{pmatrix} 0.029 & 0.002 \\ 0.002 & 0.030 \end{pmatrix}, \begin{pmatrix} 0.090 & 0.002 \\ 0.002 & 0.090 \end{pmatrix}$
$p(z,0)$	$[0.021, 0.173, 0.81]$

TABLE C.2: Parameters learned from 2D, three-well SDE data

D

APPENDIX D: NONPARAMETRIC INFERENCE FOR CONFORMATIONAL SWITCHING

D.1 VARIATIONAL MESSAGE PASSING

To derive the variational message passing algorithm, first recall that the full sequence distribution reads

$$q(z_{[1,T]}) \propto \exp\{\mathsf{E}[\ln p(z_1) + \ln p(x_1 \mid \theta, z_1) \\ + \sum_{t=2}^{T} \ln p(z_t \mid \Pi_{z_{t-1}}) + \ln p(x_t \mid \Theta_{z_t})]\}. \qquad \text{(D.1.1)}$$

Define

$$\alpha(z,1) := \exp\{\mathsf{E}[\ln p(z,1)] + \mathsf{E}[\ln p(x(1) \mid \Theta_z)]\}$$

to find

$$\begin{aligned} q(z_2,2) \propto \sum_{z_1} \alpha(z_1,1) \exp\left\{\mathsf{E}\left[\ln p(z_2 \mid \Pi_{z_1}) + \ln p(x(2) \mid \Theta_{z_2})\right]\right\} \\ \sum_{z_3,\ldots,z_T} \exp\left\{\mathsf{E}\left[\sum_{t=3} \ln p(z_t \mid \Pi_{z_{t-1}}) + \ln p(x(t) \mid \Theta_{z_t})\right]\right\} \\ \propto \exp\left\{\mathsf{E}\left[\ln p(x(2) \mid \Theta_{z_2})\right]\right\} \sum_{z_1} \alpha(z_1,1) \exp\{\ln p(z_2 \mid \Pi_{z_1})\} \\ \sum_{z_3,\ldots,z_T} \exp\left\{\sum_{t=3} \ln p(z_t \mid \Pi_{z_{t-1}}) + \ln p(x(t) \mid \Theta_{z_t})\right\}. \end{aligned} \qquad \text{(D.1.2)}$$

Accordingly, by defining the forward-messages

$$\alpha(z,t) := \exp\{\mathsf{E}[\ln p(x(t) \mid \Theta_z)]\} \sum_{z'} \alpha(z', t-1) \exp\{\mathsf{E}[\ln p(z \mid \Pi_{z'})]\} \tag{D.1.3}$$

this yields

$$q(z,t) \propto \alpha(z,t) \sum_{z_{t+1},\ldots,z_T} \exp\left\{ \sum_{t=3} \ln p(z_t \mid \Pi_{z_{t-1}}) + \ln p(x(t) \mid \Theta_{z_t}) \right\}. \tag{D.1.4}$$

Repeating this procedure from the last time step T backwards with the backward-messages

$$\beta(z,t) := \sum_{z_{t+1}} \exp\{\mathsf{E}[\ln p(x(t+1) \mid \Theta_{z_{t+1}})]\} \beta(z_{t+1}, t+1) \exp\{\mathsf{E}[\ln p(z_{t+1} \mid \Pi_{z_t})]\}$$

with initial value $\beta(z,T) = 1 \forall z$, one can compute the marginals as $q(z,t) \propto \alpha(z,t)\beta(z,t)$.

The involved expectations can all be evaluated in closed form. First,

$$\begin{aligned} \mathsf{E}[\ln p(z \mid \Pi_{z'})] &= \mathsf{E}[\ln \Pi_{z',z}] \\ &= \psi(\kappa\beta_z + \xi\delta_{z'z}) - \psi\left(\xi + \kappa \sum_k \beta_k\right), \end{aligned} \tag{D.1.5}$$

with the digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x)$.

The expectations with respect to Θ

$$\mathsf{E}[\ln p(x \mid \Theta_z)] = -n \ln\left(\sqrt{2\pi}\right) - \frac{1}{2} \mathsf{E}\left[\ln |\Sigma_z|\right] - \frac{1}{2} \mathsf{E}[(x-\mu_z)^\top \Sigma_z^{-1} (x - \mu_z)], \tag{D.1.6}$$

where $x \in \mathbb{R}^n$. Notice that for readability, we did not introduce additional symbols to denote the random variables of which μ, Σ are realizations of; we use the same symbol for both and the expectation is to be taken over both of these random variables.

The expectation of the log-determinant of an IW distributed quantity is known to be [34]

$$\begin{aligned} \mathsf{E}[\ln |\Sigma|] &= \psi_n\left(\frac{\nu}{2}\right) + n \ln(2) + \ln |\Psi^{-1}| \\ &= \sum_{i=1}^n \psi\left(\frac{\nu - i + 1}{2}\right) + n \ln(2) + \ln |\Psi^{-1}|. \end{aligned} \tag{D.1.7}$$

Secondly, one can straightforwardly compute

$$\begin{aligned} \mathsf{E}[(x-\mu_z)^\top \Sigma_z^{-1} (x-\mu_z)] &= \mathsf{E}_{\mu_z}[(x-\mu_z)^\top \mathsf{E}_{\Sigma_z}[\Sigma_z^{-1}](x-\mu_z)] \\ &= 2x^\top \mathsf{E}_{\Sigma_z}[\Sigma_z^{-1}]x - x^\top \mathsf{E}_{\Sigma_z}[\Sigma_z^{-1}] \mathsf{E}_{\mu_z}[\mu_z] \\ &- 2\,\mathsf{E}_{\mu_z}[\mu_z^\top] \mathsf{E}_{\Sigma_z}[\Sigma_z^{-1}]x + \mathsf{E}_{\mu_z,\Sigma_z}[\mu_z^\top \Sigma_z^{-1} \mu_z] \end{aligned} \tag{D.1.8}$$

The means are normally distributed, $\mu_z \sim \mathcal{N}(\mu_{0,z}, \Sigma_z/\lambda_z)$, hence $\mathsf{E}_{\mu_z}[\mu_z] = \mu_{0,z}$. The expectation with respect to Σ_z is the expectation of the inverse of an IW distributed random variable; consequently, it is just the expectation of the Wishart distributed random variable $\Sigma_z^{-1} \sim \mathrm{W}(\Psi_z^{-1}, \nu_z)$: $\mathsf{E}_{\Sigma_z}[\Sigma_z^{-1}] = \nu_z \Psi_z^{-1}$. We get

$$\begin{aligned} \mathsf{E}[(x-\mu_z)^\top \Sigma_z^{-1}(x-\mu_z)] &= \nu_z x^\top \Psi_z^{-1} x - \nu_z x^\top \Psi_z^{-1} \mu_{0,z} \\ &\quad - \nu_z \mu_{0,z}^\top \Psi_z^{-1} x + \mathsf{E}_{\mu_z,\Sigma_z}[\mu_z^\top \Sigma_z^{-1} \mu_z] \\ &= \nu_z x^\top \Psi_z^{-1} x - \nu_z x^\top \Psi_z^{-1} \mu_{0,z} \\ &\quad - \nu_z \mu_{0,z}^\top \Psi_z^{-1} x + \left(\frac{n}{\lambda_z} + \nu_z \mu_{0,z}^\top \Psi_z^{-1} \mu_{0,z} \right) \end{aligned} \tag{D.1.9}$$

where an identity for quadratic forms was used: if $x \in \mathbb{R}^n$ is a random variable and $\mathrm{C} \in \mathbb{R}^{n\times n}$ is symmetric, $C_{ij} = C_{ji}$, then $x^\top C x$ is called a *quadratic form* and

$$\mathsf{E}[x^\top C x] = \mathrm{tr}(C\Sigma) + \mu^\top C \mu \tag{D.1.10}$$

where μ, Σ are the mean and covariance of x. Note also that due to symmetry $\mu^\top \Psi^{-1} x = \left(x^\top \Psi^{-1} \mu\right)^\top = x^\top \Psi^{-1} \mu$. This completes the computation of the expected log-likelihood Eq. (D.1.6) and hence, the variational forward and backward messages.

D.2 ESTIMATION OF THE TOP-LEVEL STICK-BREAKING MEASURE

Because of non-conjugacy between $p(\beta)$ and $p(\pi \mid \beta)$, one cannot readily write down a closed-form variational posterior $q(\beta)$. Is is customary to instead utilize a point estimate, which we obtain by gradient optimization of the ELBO with respect to β. The following derivation follows the one presented in [158] and is reiterated here for completeness. Consider

$$\begin{aligned} \partial_\beta \mathsf{L} &= \partial_\beta \, \mathsf{E}\left[\ln p(\beta, \Pi)\right] \\ &= \partial_\beta \left\{ \ln p(\beta) + \sum_{z=1}^{K} \mathsf{E}[\ln p(\Pi_z \mid \beta)] \right\}, \end{aligned} \tag{D.2.11}$$

where the expectation is taken with respect to the variational distributions $q(\pi_z \mid \eta_{z,1}, ..., \eta_{z,K+1})$. For any partition of the domain, we have - by definition of the DP -

$$p(\pi_z \mid \beta) = \mathrm{Dir}(\kappa\beta_1, \kappa\beta_2, \ldots, \kappa\beta_z + \xi, \ldots, \kappa\beta_K, \kappa\beta_{K+1}).$$

where β_{K+1} denotes the “rest” of the stick-breaking measure, $\beta_{K+1} = 1 - \sum_{k=1}^{K} \beta_k$. Therefore,

$$\mathsf{E}\left[\ln p(\Pi_z \mid \beta, \kappa, \xi)\right] = \ln \frac{\Gamma\left(\kappa \sum_{z'} \left(\beta_{z'} + \frac{\xi}{\kappa}\delta_{z,z'}\right)\right)}{\prod_{z'} \Gamma(\kappa\beta_{z'} + \xi\delta_{z,z'})} + \sum_{z'=1}^{K} \left(\kappa\beta_{z'} + \xi\delta_{z,z'} - 1\right) \mathsf{E}[\ln \Pi_{z,z'}],$$

where the latter expectation is known to be

$$\mathsf{E}[\ln \pi_{z,z'}] = \psi(\eta_{z,z'}) - \psi\left(\sum_{z''} \eta_{z,z''}\right).$$

Notice that

$$\kappa\left(\sum_{z=1}^{K+1}\beta_z + \frac{\xi}{\kappa}\delta_{z',z}\right) = \kappa + \xi.$$

With this,

$$\begin{aligned}\partial_{\beta_k}\,\mathsf{E}[\ln p(\Pi_z \mid \beta)] = \partial_{\beta_k}&\Bigg[\ln\Gamma(\kappa+\xi) - \sum_{j=1}^{K+1}\ln\Gamma\left(\kappa\beta_j + \xi\delta_{z,j}\right)\\ &\quad + \sum_{j=1}^{K+1}\left(\beta\sigma_j + \xi\delta_{z,j} - 1\right)\left(\psi(\eta_{z,j}) - \psi\left(\sum_j \eta_{z,j}\right)\right)\Bigg]\\ &= -\kappa\psi\left(\kappa\beta_k + \xi\delta_{z,k}\right) + \kappa\psi(\kappa\beta_{K+1}) + \kappa\psi(\eta_{z,k}) - \kappa\psi\left(\sum_j \eta_{z,j}\right)\\ &\quad - \kappa\left(\psi(\eta_{z,K+1}) - \psi\left(\sum_j \eta_{z,j}\right)\right)\\ &= \kappa\left(-\psi\left(\kappa\beta_k + \xi\delta_{z,k}\right) + \psi(\kappa\beta_{K+1}) + \psi(\eta_{z,k}) - \psi(\eta_{z,K+1})\right).\end{aligned}$$

Summation yields

$$\begin{aligned}\sum_{z=1}^{K}\partial_{\beta_k}\,\mathsf{E}[\ln p(\Pi_z \mid \beta,\kappa,\xi)] = \kappa&\Bigg(\sum_{z=1}^{K}\left[-\psi\left(\kappa\beta_k + \xi\delta_{z,k}\right) + \psi(\eta_{z,k}) - \psi(\eta_{z,K+1})\right]\\ &\quad + K\psi(\kappa\beta_{K+1})\Bigg).\end{aligned} \tag{D.2.12}$$

The gradient of the log-prior can be computed by back-transforming the stick-breaking measure: given the stick-breaking variables $\beta_i = \epsilon_i \prod_{j<i}(1-\epsilon_j)$, we can invert this relation to arrive at

$$\epsilon_i(\beta) = \frac{\beta_i}{1 - \sum_{j<i}\beta_j}.$$

Applying the rule for transformation of probability densities one obtains

$$\ln p(\beta) = \ln p_\epsilon(\epsilon) + \ln\left|\frac{\partial\epsilon}{\partial\beta}\right| \tag{D.2.13}$$

with

$$\ln p_\epsilon(\epsilon) = \sum_{k=1}^{K+1}\left((\gamma-1)\ln(1-\epsilon_k) - \mathrm{B}(1,\gamma)\right)$$

and the normalizer of the Beta distribution, $\mathrm{B}(1,\gamma) = \Gamma(1)\Gamma(\gamma)/\Gamma(1+\gamma)$. The Jacobian determinant can be evaluated by noting that

$$\left(\frac{\partial\epsilon}{\partial\beta}\right)_{ij} = \begin{cases} 0 & \text{if } i<j \\ \frac{1}{1-\sum_{k<i}\beta_k} & \text{if } i=j \\ \frac{-\beta_i}{(1-\sum_{k<i}\beta_k)^2} & \text{if } i>j \end{cases}. \tag{D.2.14}$$

Hence,

$$\partial_{\beta_z} \ln p(\beta) = 2 \sum_{i \geq z} \ln \frac{1}{1 - \sum_{j<i} \beta_j} - (\gamma - 1) \sum_{i \geq z} \ln \frac{1}{1 - \sum_{j \leq i} \beta_j}, \tag{D.2.15}$$

and Eq. (D.2.11) is given by summation of Eqs. (D.2.12) and (D.2.15).

D.3 EXPERIMENTAL DETAILS

D.3.1 *Synthetic data generation*

In the synthetic HMM examples, we generate $I = 10$ independent latent sequences $z^i_{[1,T]}$ consisting of $T = 1000$ time points each. In the three well example, we generate 10 trajectories of length $T = 10000$ time points each.

D.3.2 *Hyperparameter initialization*

The variational marginals $q(z, t)$ are initialized as

$$q(z, t) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1) \, \forall \, z \tag{D.3.16}$$

with subsequent normalization.

The parameters of the transition distributions $q(\pi_z)$ are set up by drawing one $K+1$-dimensional stick-breaking measure β and setting

$$\eta_{z,1}, \ldots, \eta_{z,K+1} = \kappa\beta + \xi\delta_z$$

for all z. For all experiments $\gamma = \kappa = 0.6$. The prior P_0 is initialized empirically, which is common in the field and which can be understood as a type of Empirical Bayes initialization [222]; specifically, we set the NIW parameters ν and λ as well as the stickiness parameters ξ to 1% of the number of data points:

$$\nu = \lambda = \xi = 0.01 \cdot IT. \tag{D.3.17}$$

This is justified by the fact that the variational updates for ν and λ are on the order of magnitude of the number of data points, see Eq. (5.4.29). This translates to the assumption that the quantities of interest are on the order of magnitude of the observed data. This is particularly intuitive in the case of stickiness, because this means that we are interested in processes with sojourn times that are observable in the data. Much smaller sojourn times could not be properly resolved and longer ones not observed at all.

The NIW scale and mean for data $x \in \mathbb{R}^n$ are initialized as

$$\begin{aligned} \Psi &= \frac{1}{IT - 1} \sum_{t=1}^{T} \sum_{i=1}^{I} (x_{i,t} - \bar{x})(x_{i,t} - \bar{x})^\top (\nu_k - n - 1), \\ \mu_0 &= \bar{x} = \frac{1}{IT} \sum_{i,t} x_{i,t}, \end{aligned} \tag{D.3.18}$$

for all z. For $x \in [0, 2\pi)^n$, on the other hand, we set

$$\begin{aligned} \Psi_k &= 0.1 \cdot \mathbb{1}^{n \times n} \cdot (\nu_z - n - 1) \\ \mu_{0,z,i} &\sim \mathrm{Uniform}(0, 2\pi) \end{aligned} \tag{D.3.19}$$

where $\mathbb{1}^{n \times n}$ is the n-dimensional identity matrix and $\mu_{0,k,i}$ denotes the i-th entry of $\mu_{0,k}$. This ensures a good coverage of the $[0, 2\pi)^n$ space with a covariance on an order of magnitude of the observable region.

Finally, to alleviate the initialization-dependency of the gradient ascent scheme, we utilize a multi-start approach: the random initialization of $q(z, t)$ and $B = \beta$ creates different initial conditions in each VI instance. We hence run several instances of the inference algorithm until convergence and then select the one with the maximal ELBO score as the overall optimum.

NOTATION

SYMBOL	DESCRIPTION
$\mathbb{N}$	The set of natural numbers.
$\mathbb{N}_{>i}, \mathbb{N}_{\geq i}$	The set of natural numbers with elements greater (or equal) than i.
$\mathbb{R}$	The set of real numbers.
$\mathbb{R}_{>t}, \mathbb{R}_{\geq t}$	The set of real numbers with elements greater (or equal) than t.
Δ^n	The n-dimensional probability simplex; i.e., $\Delta^n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1 \wedge x_j \geq 0, \forall j \in \{1, \ldots, n\}\}$.
$\mathbb{1}(\cdot)$	The indicator function.
$\mathbb{1}_{n\times n}$	The n-dimensional identity matrix.
$\delta_x(\cdot) = \delta(x - \cdot)$	The Dirac delta function/point measure at x.
$\delta_{x,x'}$	The Kronecker delta function.
$f(x)$	A function f of a variable x.
$J[f]$	A functional J of a function f.
$\partial_x(\cdot)$	The gradient with respect to x.
$\mathsf{P}(\cdot)$	A probability measure.
$p(\cdot)$	A probability density function or probability mass function.
$\mathsf{E}[\cdot]$	The expectation operator.
$\mathsf{D}_{\mathrm{KL}}[p(x) \parallel q(x)]$	The Kullback-Leibler divergence between $p(x)$ and $q(x)$.
$\mathsf{L}[q]$	The evidence lower bound dependent on the distribution q.
$\mathrm{Gam}(\cdot \mid \alpha, \beta)$	Probability density function of the gamma distribution with shape parameter α and rate parameter β.
$\mathrm{Cat}(\cdot \mid \pi)$	Probability mass function of the categorical distribution with probability vector π
$\mathrm{Exp}(\cdot \mid \lambda)$	Probability density function of the exponential distribution with rate parameter λ.
$\mathrm{Uniform}(\cdot \mid a, b)$	Probability density function of the uniform distribution with lower bound a and upper bound b.
$\mathcal{N}(\cdot \mid \mu, \Sigma)$	Probability density function of the (multivariate) normal distribution with mean μ and variance/covariance matrix Σ.
$\mathrm{Dir}(\cdot \mid \alpha)$	Probability density function of the Dirichlet distribution with concentration parameter vector α.
$\mathrm{Mult}(\cdot \mid \pi)$	Probability mass function of the multinomial distribution with probability vector π.
$\mathrm{GEM}(\gamma)$	Probability measure of the Griffiths–Engen–McCloskey process with concentration parameter γ.

ACRONYMS

BvM	bivariate von Mises
càdlàg	continue à droite, limite à gauche
CAVI	coordinate-ascent variational inference
CTMC	continuous-time Markov chain
DNA	deoxyribonucleic acid
DS	dynamical system
DTMC	discrete-time Markov chain
DP	Dirichlet process
EL	Euler-Lagrange
ELBO	evidence lower-bound
FPE	Fokker-Planck equation
FRET	fluorescence resonance energy transfer
GFP	green fluorescent protein
GLDS	Gaussian linear dynamical system
GP	Gaussian process
HDP	hierarchical Dirichlet process
HDP-HMM	hierarchical Dirichlet process hidden Markov model
HME	hybrid master equation
HMM	hidden Markov model
IW	inverse-Wishart
KBE	Kolmogorov backward equation
KFE	Kolmogorov forward equation
KL	Kullback-Leibler
LDS	linear dynamical system
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MD	molecular dynamics
MJP	Markov jump process
MN	Matrix-Normal
MSM	Markov state model
NIW	Normal-inverse-Wishart
ODE	ordinary differential equation
PCCA	Perron-cluster cluster analysis
PDE	partial differential equation
PDF	probability density function
PMF	probability mass function
RNA	ribonucleic acid
RNAP	RNA polymerase
RTS	Rauch-Tung-Striebel
SDE	stochastic differential equation
SDS	stochastic dynamical system
SLDS	switching linear dynamical system
SSDE	switching stochastic differential equation
TL	translation
TX	transcription
VEM	variational expectation maximization
VI	variational inference
vM	von Mises

BIBLIOGRAPHY

[1] L. Köhs, B. Alt, and H. Koeppl, "Markov chain Monte Carlo for continuous-time switching dynamical systems", *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, pp. 11 430–11 454, 2022.

[2] L. Köhs, B. Alt, and H. Koeppl, "Variational inference for continuous-time switching dynamical systems", *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[3] L. Köhs, K. Kukovetz, O. Rauh, and H. Koeppl, "Nonparametric Bayesian inference for meta-stable conformational dynamics", *Physical Biology*, vol. 19, no. 5, p. 056 006, 2022.

[4] P. C. Bressloff, *Stochastic processes in cell biology*. Springer, 2014, vol. 41.

[5] O. Kallenberg and O. Kallenberg, *Foundations of modern probability*. Springer, 1997.

[6] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*. Springer Science & Business Media, 2013.

[7] J. R. Norris, *Markov Chains*. Cambridge University Press, 1997.

[8] R. C. Robinson, *An introduction to dynamical systems: continuous and discrete*. American Mathematical Soc., 2012.

[9] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[11] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[12] O. L. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

[13] P. Del Moral and S. Penev, *Stochastic Processes: From Applications to Theory*. Chapman and Hall/CRC, 2017.

[14] S. N. Ethier and T. G. Kurtz, *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[15] B. Øksendal, *Stochastic differential equations*. Springer, 2003.

[16] C. Kipnis and C. Landim, *Scaling limits of interacting particle systems*. Springer Science & Business Media, 1998, vol. 320.

[17] J. L. Doob, "Markoff chains–denumerable case", *Transactions of the American Mathematical Society*, vol. 58, no. 3, pp. 455–473, 1945.

[18] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions", *Journal of computational physics*, vol. 22, no. 4, pp. 403–434, 1976.

[19] P. W. Lewis and G. S. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning", *Naval research logistics quarterly*, vol. 26, no. 3, pp. 403–413, 1979.

[20] S. Särkkä and A. Solin, *Applied stochastic differential equations*. Cambridge University Press, 2019.

[21] S. Särkkä, *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

[22] F. B. Hanson, *Applied stochastic processes and control for jump-diffusions: modeling, analysis and computation*. SIAM, 2007.

[23] T. Sutter, A. Ganguly, and H. Koeppl, "A variational approach to path estimation and parameter inference of hidden diffusion processes", *Journal of Machine Learning Research*, vol. 17, no. 190, pp. 1–37, 2016.

[24] T. Sottinen and S. Särkkä, "Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems", *Bayesian Analysis*, vol. 3, no. 3, pp. 555–584, 2008.

[25] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

[26] X. Mao and C. Yuan, *Stochastic differential equations with Markovian switching*. Imperial college press, 2006.

[27] C. G. Cassandras and S. Lafortune, *Introduction to discrete event systems*. Springer Science & Business Media, 2009.

[28] P. S. Maybeck, *Stochastic models, estimation, and control*. Academic press, 1982.

[29] F. E. Daum, "Exact finite dimensional nonlinear filters for continuous time processes with discrete time measurements", in *The 23rd IEEE Conference on Decision and Control*, IEEE, 1984, pp. 16–22.

[30] A. Bain and D. Crisan, *Fundamentals of stochastic filtering*. Springer Science & Business Media, 2008, vol. 60.

[31] B. D. O. Anderson and I. B. Rhodes, "Smoothing algorithms for nonlinear finite-dimensional systems", *Stochastics: An International Journal of Probability and Stochastic Processes*, vol. 9, no. 1-2, pp. 139–165, 1983.

[32] L. Huang, L. Pauleve, C. Zechner, M. Unger, A. S. Hansen, and H. Koeppl, "Reconstructing dynamic molecular states from single-cell time series", *Journal of The Royal Society Interface*, vol. 13, no. 122, p. 20 160 533, 2016.

[33] P. Orbanz, "Construction of nonparametric bayesian models from parametric bayes equations", *Advances in neural information processing systems*, vol. 22, 2009.

[34] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[35] S. Ghosal and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

[36] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems", *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[37] J. Sethuraman, "A constructive definition of dirichlet priors", *Statistica sinica*, pp. 639–650, 1994.

[38] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet Processes", *Journal of the American Statistical Association*, vol. 101, no. 476, 2006.

[39] R. Bellman, “Dynamic programming”, *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[40] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering”, *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[41] V. Rao and Y. W. Teh, “Fast MCMC sampling for Markov jump processes and extensions.”, *Journal of Machine Learning Research*, vol. 14, no. 11, 2013.

[42] M. Opper and G. Sanguinetti, “Variational inference for Markov jump processes”, *Advances in neural information processing systems*, vol. 20, pp. 1105–1112, 2007.

[43] C. Wildner and H. Koeppl, “Moment-based variational inference for stochastic differential equations”, in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 130, PMLR, 2021, pp. 1918–1926.

[44] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, “An introduction to MCMC for machine learning”, *Machine learning*, vol. 50, no. 1, pp. 5–43, 2003.

[45] C. K. Carter and R. Kohn, “Markov chain monte carlo in conditionally gaussian state space models”, *Biometrika*, vol. 83, no. 3, pp. 589–601, 1996.

[46] G. L. Jones and Q. Qin, “Markov chain monte carlo in practice”, *Annual Review of Statistics and Its Application*, vol. 9, pp. 557–578, 2022.

[47] D. Gamerman and H. F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press, 2006.

[48] J. Besag, “Spatial interaction and the statistical analysis of lattice systems”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 192–225, 1974.

[49] H.-C. Ruiz and H. J. Kappen, “Particle smoothing for hidden diffusion processes: Adaptive path integral smoother”, *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 3191–3203, 2017.

[50] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians”, *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[51] M. Johnson and A. Willsky, “Stochastic variational inference for bayesian time series models”, in *International Conference on Machine Learning*, PMLR, 2014, pp. 1854–1862.

[52] Z. Dong, B. Seybold, K. Murphy, and H. Bui, “Collapsed amortized variational inference for switching nonlinear dynamical systems”, in *International Conference on Machine Learning*, PMLR, 2020, pp. 2638–2647.

[53] A. Zhang, S. Gultekin, and J. Paisley, “Stochastic variational inference for the hdp-hmm”, in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 51, 2016, pp. 800–808.

[54] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom), 2003.

[55] R. Van Handel, “Filtering, stability, and robustness”, Ph.D. dissertation, California Institute of Technology, 2007.

[56] S. K. Mitter and N. J. Newton, "A variational approach to nonlinear estimation", *SIAM journal on control and optimization*, vol. 42, no. 5, pp. 1813–1833, 2003.

[57] A. G. d. G. Matthews, J. Hensman, R. Turner, and Z. Ghahramani, "On sparse variational methods and the Kullback-Leibler divergence between stochastic processes", in *Artificial Intelligence and Statistics*, PMLR, 2016, pp. 231–239.

[58] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv", *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983.

[59] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton university press, 2011.

[60] Y. Weng, X. Liu, H. Hu, H. Huang, S. Zheng, Q. Chen, J. Song, B. Cao, J. Wang, S. Wang, *et al.*, "Open eyes and closed eyes elicit different temporal properties of brain functional networks", *NeuroImage*, vol. 222, p. 117 230, 2020.

[61] J. Taghia, W. Cai, S. Ryali, J. Kochalka, J. Nicholas, T. Chen, and V. Menon, "Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition", *Nature communications*, vol. 9, no. 1, pp. 1–19, 2018.

[62] M. L. Kringelbach and G. Deco, "Brain states and transitions: Insights from computational neuroscience", *Cell Reports*, vol. 32, no. 10, p. 108 128, 2020.

[63] G. Deco, J. Cruzat, J. Cabral, E. Tagliazucchi, H. Laufs, N. K. Logothetis, and M. L. Kringelbach, "Awakening: Predicting external stimulation to force transitions between different brain states", *Proceedings of the National Academy of Sciences*, vol. 116, no. 36, pp. 18 088–18 097, 2019.

[64] T. Tian and K. Burrage, "Stochastic models for regulatory networks of the genetic toggle switch", *Proceedings of the national Academy of Sciences*, vol. 103, no. 22, pp. 8372–8377, 2006.

[65] D. F. Anderson, B. Ermentrout, and P. J. Thomas, "Stochastic representations of ion channel kinetics and exact stochastic simulation of neuronal dynamics", *Journal of computational neuroscience*, vol. 38, no. 1, pp. 67–82, 2015.

[66] M. A. Rapsomaniki, S. Maxouri, P. Nathanailidou, M. R. Garrastacho, N. N. Giakoumakis, S. Taraviras, J. Lygeros, and Z. Lygerou, "In silico analysis of DNA re-replication across a complete genome reveals cell-to-cell heterogeneity and genome plasticity", *NAR genomics and bioinformatics*, vol. 3, no. 1, lqaa112, 2021.

[67] J. Lygeros, K. Koutroumpas, S. Dimopoulos, I. Legouras, P. Kouretas, C. Heichinger, P. Nurse, and Z. Lygerou, "Stochastic hybrid modeling of dna replication across a complete genome", *Proceedings of the National Academy of Sciences*, vol. 105, no. 34, pp. 12 295–12 300, 2008.

[68] P. C. Bressloff, "Stochastic switching in biology: From genotype to phenotype", *Journal of Physics A: Mathematical and Theoretical*, vol. 50, no. 13, p. 133 001, 2017.

[69] J. Lygeros and M. Prandini, "Stochastic hybrid systems: A powerful framework for complex, large scale applications", *European Journal of Control*, vol. 16, no. 6, pp. 583–594, 2010.

[70] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems", *Automatica*, vol. 12, no. 6, pp. 601–611, 1976.

[71] M. Střelec, K. Macek, and A. Abate, "Modeling and simulation of a microgrid as a stochastic hybrid system", in *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, IEEE, 2012, pp. 1–9.

[72] S. K. Khaitan and J. D. McCalley, "Design techniques and applications of cyberphysical systems: A survey", *IEEE Systems Journal*, vol. 9, no. 2, pp. 350–365, 2014.

[73] R. G. Sanfelice *et al.*, "Analysis and design of cyber-physical systems. a hybrid control systems approach", in *Cyber-physical systems: From theory to practice*, CRC Press Boca Raton, FL, USA, 2016, pp. 1–29.

[74] M. Azzouzi and I. T. Nabney, "Modelling financial time series with switching state space models", in *Proceedings of the IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, IEEE, 1999, pp. 240–249.

[75] C. M. Carvalho and H. F. Lopes, "Simulation-based sequential analysis of Markov switching stochastic volatility models", *Computational Statistics & Data Analysis*, vol. 51, no. 9, pp. 4526–4542, 2007.

[76] T. Duprey and B. Klaus, "How to predict financial stress? an assessment of markov switching models", *ECB Working Paper Series*, 2017.

[77] K. Salhi, M. Deaconu, A. Lejay, N. Champagnat, and N. Navet, "Regime switching model for financial data: Empirical risk analysis", *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 148–157, 2016.

[78] M. H. Davis, "Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 46, no. 3, pp. 353–376, 1984.

[79] J. Hu, J. Lygeros, and S. Sastry, "Towards a theory of stochastic hybrid systems", in *International Workshop on Hybrid Systems: Computation and Control*, Springer, 2000, pp. 160–173.

[80] S. Engell, G. Frehse, and E. Schnieder, *Modelling, analysis and design of hybrid systems*. Springer, 2003, vol. 279.

[81] R. Goebel, R. G. Sanfelice, and A. R. Teel, "Hybrid dynamical systems", in *Hybrid dynamical systems*, Princeton University Press, 2012.

[82] M. H. A. Davis, *Markov Models and Optimization*. Routledge, 2018.

[83] C. G. Cassandras and J. Lygeros, *Stochastic hybrid systems*. CRC Press, 2018.

[84] A. Golightly and D. J. Wilkinson, "Bayesian inference for nonlinear multivariate diffusion models observed with error", *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1674–1693, 2008.

[85] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead, "Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion)", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 333–382, 2006.

[86] V. Stathopoulos and M. A. Girolami, “Markov chain monte carlo inference for markov jump processes via the linear noise approximation”, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20 110 541, 2013.

[87] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, “Gaussian process approximations of stochastic differential equations”, *Journal of Machine Learning Research*, vol. 1, pp. 1–16, 2007.

[88] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor, “Variational inference for diffusion processes”, *Advances in Neural Information Processing Systems*, vol. 20, pp. 17–24, 2007.

[89] C. Wildner and H. Koeppl, “Moment-based variational inference for Markov jump processes”, in *International Conference on Machine Learning*, PMLR, 2019, pp. 6766–6775.

[90] D. Alspach and H. Sorenson, “Nonlinear bayesian estimation using gaussian sum approximations”, *IEEE transactions on automatic control*, vol. 17, no. 4, pp. 439–448, 1972.

[91] G. Böker and J. Lunze, “Stability and performance of switching Kalman filters”, *International Journal of Control*, vol. 75, no. 16-17, pp. 1269–1281, 2002.

[92] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams, “Composing graphical models with neural networks for structured representations and fast inference”, *Advances in Neural Information Processing Systems 29*, pp. 2954–2962, 2016.

[93] S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski, “Bayesian learning and inference in recurrent switching linear dynamical systems”, in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 914–922.

[94] J. Glaser, M. Whiteway, J. P. Cunningham, L. Paninski, and S. Linderman, “Recurrent switching dynamical systems models for multiple interacting neural populations”, in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020.

[95] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert, “Data-driven MCMC for learning and inference in switching linear dynamic systems”, Georgia Institute of Technology, 2005.

[96] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, “Nonparametric Bayesian learning of switching linear dynamical systems”, *Advances in Neural Information Processing Systems*, vol. 21, pp. 457–464, 2008.

[97] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta, “Composing graphical models with neural networks for structured representations and fast inference”, *Advances in neural information processing systems*, vol. 29, pp. 2946–2954, 2016.

[98] M. Opper, A. Ruttor, and G. Sanguinetti, “Approximate inference in continuous time gaussian-jump processes”, *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[99] F. Stimberg, M. Opper, G. Sanguinetti, and A. Ruttor, “Inference in continuous-time change-point models”, *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[100] F. Stimberg, A. Ruttor, and M. Opper, "Bayesian inference for change points in dynamical systems with reusable states-a chinese restaurant process approach", in *Artificial Intelligence and Statistics*, PMLR, 2012, pp. 1117–1124.

[101] E. Çınlar, *Probability and Stochastics*. Springer New York, 2011.

[102] R. F. Pawula, "Generalizations and extensions of the Fokker-Planck-Kolmogorov equations", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 33–41, 1967.

[103] C. Grossmann, H.-G. Roos, and M. Stynes, *Numerical treatment of partial differential equations*. Springer, 2007.

[104] E. Pardoux, "Stochastic partial differential equations and filtering of diffusion processes", *Stochastics*, vol. 3, no. 1-4, pp. 127–167, 1980.

[105] A. Samoilenko and M. Perestyuk, *Impulsive differential equations*. World scientific, 1995.

[106] C. Archambeau and M. Opper, "Approximate inference for continuous-time Markov processes", *Bayesian Time Series Models*, pp. 125–140, 2011.

[107] T. L. Guo and K. Zhang, "Impulsive fractional partial differential equations", *Applied Mathematics and Computation*, vol. 257, pp. 581–590, 2015.

[108] J. Besag, P. Green, D. Higdon, and K. Mengersen, "Bayesian computation and stochastic systems", *Statistical science*, pp. 3–41, 1995.

[109] M. Mider, M. Schauer, and F. van der Meulen, "Continuous-discrete smoothing of diffusions", *Electronic Journal of Statistics*, vol. 15, no. 2, pp. 4295–4342, 2021.

[110] L. C. Evans, *Partial differential equations*. American Mathematical Soc., 2010, vol. 19.

[111] R. F. Stengel, *Optimal control and estimation*. Courier Corporation, 1994.

[112] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff problems*. Springer, 1993.

[113] R. L. Stratonovich, "Conditional markov processes and their application to the theory of optimal control", 1968.

[114] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering",*Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, vol. 2, no. 3, pp. 347–369, 1964.

[115] B. D. Anderson, "Reverse-time diffusion equation models", *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[116] M. Davis, "Pathwise solutions and multiplicative functionals in nonlinear filtering", in *1979 18th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes*, IEEE, vol. 2, 1979, pp. 176–181.

[117] D. Crisan, J. Diehl, P. K. Friz, and H. Oberhauser, "Robust filtering: Correlated noise and multidimensional observation", *The Annals of Applied Probability*, vol. 23, no. 5, pp. 2139–2160, 2013.

[118] X. Mao, "Stability of stochastic differential equations with Markovian switching", *Stochastic processes and their applications*, vol. 79, no. 1, pp. 45–67, 1999.

[119] E. B. Fox, "Bayesian nonparametric learning of complex dynamical phenomena", Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[120] N. Shephard and M. K. Pitt, "Likelihood analysis of non-Gaussian measurement time series", *Biometrika*, vol. 84, no. 3, pp. 653–667, 1997.

[121] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin diffusions",*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 1, pp. 255–268, 1998.

[122] A. Hofmann, J. Falk, T. Prangemeier, D. Happel, A. Köber, A. Christmann, H. Koeppl, and H. Kolmar, "A tightly regulated and adjustable crispr-dcas9 based and gate in yeast", *Nucleic acids research*, vol. 47, no. 1, pp. 509–520, 2019.

[123] A. Ocone, A. J. Millar, and G. Sanguinetti, "Hybrid regulatory models: A statistically tractable approach to model regulatory network dynamics", *Bioinformatics*, vol. 29, no. 7, pp. 910–916, 2013.

[124] W. E and E. Vanden-Eijnden, "Metastability, conformation dynamics, and transition pathways in complex systems", in *Multiscale modelling and simulation*, Springer, 2004, pp. 35–68.

[125] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[126] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton university press, 2011.

[127] D. P. Bertsekas, "Nonlinear programming",*Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[128] M. McAsey, L. Mou, and W. Han, "Convergence of the forward-backward sweep method in optimal control", *Computational Optimization and Applications*, vol. 53, no. 1, pp. 207–226, 2012.

[129] D. Barber, A. T. Cemgil, and S. Chiappa, *Bayesian time series models*. Cambridge University Press, 2011.

[130] J. Sponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, *et al.*, "RNA structural dynamics as captured by molecular simulations: A comprehensive overview", *Chemical reviews*, vol. 118, no. 8, pp. 4177–4338, 2018.

[131] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation", *The Journal of Chemical Physics*, vol. 134, no. 17, p. 174 105, 2011.

[132] H. Wu, A. Mardt, L. Pasquali, and F. Noe, "Deep generative Markov state models", *Advances in Neural Information Processing Systems*, vol. 31, pp. 3975–3984, 2018.

[133] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, "Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules", *The Journal of chemical physics*, vol. 139, no. 18, 2013.

[134] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, "Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations", *The Journal of chemical physics*, vol. 146, no. 15, p. 154 104, 2017.

[135] F. Nüske, H. Wu, J.-H. Prinz, C. Wehmeyer, C. Clementi, and F. Noé, “Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias”, *The Journal of Chemical Physics*, vol. 146, no. 9, p. 094 104, 2017.

[136] F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, “Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules”, *The Journal of Chemical Physics*, vol. 139, no. 18, p. 184 114, 2013.

[137] R. J. Townshend, S. Eismann, A. M. Watkins, R. Rangan, M. Karelina, R. Das, and R. O. Dror, “Geometric deep learning of rna structure”, *Science*, vol. 373, no. 6558, pp. 1047–1051, 2021.

[138] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, *et al.*, “Highly accurate protein structure prediction with alphafold”, *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[139] G. Wei, W. Xi, R. Nussinov, and B. Ma, “Protein ensembles: How does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell”, *Chemical reviews*, vol. 116, no. 11, pp. 6516–6551, 2016.

[140] B. Sakmann and E. Neher, “Patch clamp techniques for studying ionic channels in excitable membranes”, *Annual review of physiology*, vol. 46, no. 1, pp. 455–472, 1984.

[141] C.-C. Chen, C. Cang, S. Fenske, E. Butz, Y.-K. Chao, M. Biel, D. Ren, C. Wahl-Schott, and C. Grimm, “Patch-clamp technique to characterize ion channels in enlarged individual endolysosomes”, *Nature Protocols*, vol. 12, no. 8, pp. 1639–1658, 2017.

[142] L.-M. Winterstein, K. Kukovetz, O. Rauh, D. L. Turman, C. Braun, A. Moroni, I. Schroeder, and G. Thiel, “Reconstitution and functional characterization of ion channels from nanodiscs in lipid bilayers”, *Journal of General Physiology*, vol. 150, no. 4, pp. 637–646, 2018.

[143] X. Zhuang, T. Ha, H. D. Kim, T. Centner, S. Labeit, and S. Chu, “Fluorescence quenching: A tool for single-molecule protein-folding study”, *Proceedings of the National Academy of Sciences*, vol. 97, no. 26, pp. 14 241–14 244, 2000.

[144] D. Kajihara, R. Abe, I. Iijima, C. Komiyama, M. Sisido, and T. Hohsaka, “FRET analysis of protein conformational change through position-specific incorporation of fluorescent amino acids”, *Nature Methods*, vol. 3, no. 11, pp. 923–929, 2006.

[145] B. Leimkuhler and C. Matthews, *Molecular Dynamics.* Springer, 2016.

[146] C. Matek, P. Šulc, F. Randisi, J. P. Doye, and A. A. Louis, “Coarse-grained modelling of supercoiled rna”, *The Journal of chemical physics*, vol. 143, no. 24, p. 243 122, 2015.

[147] K. A. Dill and J. L. MacCallum, “The protein-folding problem, 50 years on”, *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[148] J. Sponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, *et al.*, “RNA structural dynamics as captured by molecular simulations: A comprehensive overview”, *Chemical Reviews*, vol. 118, no. 8, pp. 4177–4338, 2018.

[149] S. Chen, R. P. Wiewiora, F. Meng, N. Babault, A. Ma, W. Yu, K. Qian, H. Hu, H. Zou, J. Wang, *et al.*, “The dynamic conformational landscape of the protein methyltransferase setd8”, *Elife*, vol. 8, e45403, 2019.

[150] V. Carnevale, L. Delemotte, and R. J. Howard, "Molecular dynamics simulations of ion channels", *Trends in Biochemical Sciences*, vol. 46, no. 7, pp. 621–622, 2021.

[151] W. Huisinga, S. Meyn, and C. Schütte, "Phase transitions and metastability in markovian and molecular systems", *The Annals of Applied Probability*, vol. 14, no. 1, pp. 419–458, 2004.

[152] K. E. Hines, J. R. Bankston, and R. W. Aldrich, "Analyzing single-molecule time series via nonparametric Bayesian inference", *Biophysical Journal*, vol. 108, no. 3, pp. 540–556, 2015.

[153] I. Sgouralis, M. Whitmore, L. Lapidus, M. J. Comstock, and S. Pressé, "Single molecule force spectroscopy at high data acquisition: A Bayesian nonparametric analysis", *The Journal of Chemical Physics*, vol. 148, no. 12, p. 123 320, 2018.

[154] C. P. Calderon and K. Bloom, "Inferring latent states and refining force estimates via hierarchical dirichlet process modeling in single particle tracking experiments", *PloS one*, vol. 10, no. 9, e0137633, 2015.

[155] B. J. Coscia, C. P. Calderon, and M. R. Shirts, "Statistical inference of transport mechanisms and long time scale behavior from time series of solute trajectories in nanostructured membranes", *The Journal of Physical Chemistry B*, vol. 124, no. 37, pp. 8110–8123, 2020.

[156] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence", in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 312–319.

[157] J. Van Gael, Y. Saatci, Y. W. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model", in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 1088–1095.

[158] M. J. Johnson *et al.*, "Bayesian time series models and scalable inference", Ph.D. dissertation, Massachusetts Institute of Technology, 2014.

[159] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models",*Journal of Chemical Theory and Computation*, vol. 11, no. 11, pp. 5525–5542, Nov. 2015.

[160] S. Cao, A. Montoya-Castillo, W. Wang, T. E. Markland, and X. Huang, "On the advantages of exploiting memory in markov state models for biomolecular dynamics", *The Journal of Chemical Physics*, vol. 153, no. 1, p. 014 105, 2020.

[161] C. Schütte and M. Sarich, "A critical appraisal of markov state models", *The European Physical Journal Special Topics*, vol. 224, no. 12, pp. 2445–2462, 2015.

[162] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, "Identification of almost invariant aggregates in reversible nearly uncoupled markov chains", *Linear Algebra and its Applications*, vol. 315, no. 1-3, pp. 39–59, 2000.

[163] S. Röblitz and M. Weber, "Fuzzy spectral clustering by pcca+: Application to markov state models and data classification", *Advances in Data Analysis and Classification*, vol. 7, no. 2, pp. 147–179, 2013.

[164] W. Wang, T. Liang, F. K. Sheong, X. Fan, and X. Huang, "An efficient bayesian kinetic lumping algorithm to identify metastable conformational states via gibbs sampling", *The Journal of Chemical Physics*, vol. 149, no. 7, p. 072 337, 2018.

[165] G. R. Bowman, "Improved coarse-graining of markov state models via explicit consideration of statistical uncertainty", *The Journal of Chemical Physics*, vol. 137, no. 13, p. 134 111, 2012.

[166] B. E. Husic and V. S. Pande, "Markov state models: From an art to a science", *Journal of the American Chemical Society*, vol. 140, no. 7, pp. 2386–2396, 2018.

[167] C. R. Schwantes, R. T. McGibbon, and V. S. Pande, "Perspective: Markov models for long-timescale biomolecular dynamics", *The Journal of chemical physics*, vol. 141, no. 9, 09B201_1, 2014.

[168] M. I. Zimmerman, K. M. Hart, C. A. Sibbald, T. E. Frederick, J. R. Jimah, C. R. Knoverek, N. H. Tolia, and G. R. Bowman, "Prediction of new stabilizing mutations based on mechanistic insights from markov state models", *ACS Central Science*, vol. 3, no. 12, pp. 1311–1321, 2017.

[169] K. A. McKiernan, B. E. Husic, and V. S. Pande, "Modeling the mechanism of cln025 beta-hairpin formation", *The Journal of Chemical Physics*, vol. 147, no. 10, p. 104 107, 2017.

[170] S. Mittal and D. Shukla, "Predicting optimal deer label positions to study protein conformational heterogeneity", *The Journal of Physical Chemistry B*, vol. 121, no. 42, pp. 9761–9770, 2017.

[171] K. M. Hart, K. E. Moeder, C. M. Ho, M. I. Zimmerman, T. E. Frederick, and G. R. Bowman, "Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators", *PloS one*, vol. 12, no. 6, e0178678, 2017.

[172] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling", *Nature Chemistry*, vol. 9, no. 10, pp. 1005–1011, 2017.

[173] B. K. Chu, M. J. Tse, R. R. Sato, and E. L. Read, "Markov state models of gene regulatory networks", *BMC Systems Biology*, vol. 11, no. 1, pp. 1–17, 2017.

[174] C. Schütte, W. Huisinga, and P. Deuflhard, "Transfer operator approach to conformational dynamics in biomolecular systems", in *Ergodic theory, analysis, and efficient simulation of dynamical systems*, Springer, 2001, pp. 191–223.

[175] D. Shukla, C. X. Hernández, J. K. Weber, and V. S. Pande, "Markov state models provide insights into dynamic modulation of protein function", *Accounts of chemical research*, vol. 48, no. 2, pp. 414–422, 2015.

[176] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, "Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations", *The Journal of Chemical Physics*, vol. 146, no. 15, p. 154 104, 2017.

[177] C. Wehmeyer and F. Noé, "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics", *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241 703, 2018.

[178] K. V. Mardia, "Bayesian analysis for bivariate von Mises distributions", *Journal of Applied Statistics*, vol. 37, no. 3, pp. 515–528, 2010.

[179] K. V. Mardia and J. Voss, "Some fundamental properties of a multivariate von Mises distribution", *Communications in Statistics - Theory and Methods*, vol. 43, no. 6, pp. 1132–1144, 2014.

[180] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck, "A generative, probabilistic model of local protein structure", *Proceedings of the National Academy of Sciences*, vol. 105, no. 26, pp. 8932–8937, 2008.

[181] A. K. Navarro, J. Frellsen, and R. E. Turner, "The multivariate generalised von mises distribution: Inference and applications", in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[182] M. J. Johnson, "Bayesian Time Series Models and Scalable Inference", Ph.D. dissertation, Massachusetts Institute of Technology, 2014.

[183] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, "The Infinite PCFG using Hierarchical Dirichlet Processes", in *Empirical Methods in Natural Language Processing*, 2007, pp. 688–697.

[184] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference.", *Journal of Machine Learning Research*, vol. 14, no. 5, 2013.

[185] M. Bryant and E. Sudderth, "Truly nonparametric online variational inference for hierarchical dirichlet processes", *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[186] M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference", Ph.D. dissertation, University of London, 2003.

[187] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, "Variational Approach to Molecular Kinetics", *Journal of Chemical Theory and Computation*, vol. 10, no. 4, pp. 1739–1752, 2014.

[188] C. R. Schwantes and V. S. Pande, "Modeling Molecular Kinetics with tICA and the Kernel Trick", *Journal of Chemical Theory and Computation*, vol. 11, no. 2, pp. 600–608, 2015.

[189] B. Plugge, S. Gazzarrini, M. Nelson, R. Cerana, J. Van, C. Derst, D. DiFrancesco, A. Moroni, G. Thiel, *et al.*, "A potassium channel protein encoded by chlorella virus PBCV-1", *Science*, vol. 287, no. 5458, pp. 1641–1644, 2000.

[190] H. Wu, F. Nüske, F. Paul, S. Klus, P. Koltai, and F. Noé, "Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations", *The Journal of Chemical Physics*, vol. 146, no. 15, p. 154 104, 2017.

[191] K. V. Mardia, C. C. Taylor, and G. K. Subramaniam, "Protein bioinformatics and mixtures of bivariate von mises distributions for angular data", *Biometrics*, vol. 63, no. 2, pp. 505–512, 2007.

[192] V. Mironov, Y. Alexeev, V. K. Mulligan, and D. G. Fedorov, "A systematic study of minima in alanine dipeptide", *Journal of Computational Chemistry*, vol. 40, no. 2, pp. 297–309, 2019.

[193] J. Grdadolnik, V. Mohacek-Grosev, R. L. Baldwin, and F. Avbelj, "Populations of the three major backbone conformations in 19 amino acid dipeptides", *Proceedings of the National Academy of Sciences*, vol. 108, no. 5, pp. 1794–1798, 2011.

[194] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations", *Journal of Molecular Biology*, vol. 7, no. 1, pp. 95–99, 1963.

[195] M. M. Sultan and V. S. Pande, "Transfer learning from Markov models leads to efficient sampling of related systems", *The Journal of Physical Chemistry B*, vol. 122, no. 21, pp. 5291–5299, 2018.

[196] M. Feig, "Is Alanine Dipeptide a Good Model for Representing the Torsional Preferences of Protein Backbones?", *Journal of Chemical Theory and Computation*, vol. 4, no. 9, pp. 1555–1564, 2008.

[197] K. Kukovetz, "Systematic analyses of structure/function variability of viral k+ channels for the development of synthetic channels", Ph.D. dissertation, Technische Universität Darmstadt, 2020.

[198] R. Schultze and S. Draber, "A nonlinear filter algorithm for the detection of jumps in patch-clamp data", *The Journal of membrane biology*, vol. 132, no. 1, pp. 41–52, 1993.

[199] M. Klaas, M. Briers, N. De Freitas, A. Doucet, S. Maskell, and D. Lang, "Fast particle smoothing: If i had a million particles", in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 481–488.

[200] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

[201] M. Betancourt and M. Girolami, "Hamiltonian monte carlo for hierarchical models", *Current trends in Bayesian methodology with applications*, vol. 79, no. 30, pp. 2–4, 2015.

[202] F. Daum, "Exact finite-dimensional nonlinear filters", *IEEE Transactions on Automatic Control*, vol. 31, no. 7, pp. 616–622, 1986.

[203] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, "The modern mathematics of deep learning", *arXiv preprint arXiv:2105.04026*, 2021.

[204] J. Berner, P. Grohs, and A. Jentzen, "Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black–scholes partial differential equations", *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 3, pp. 631–657, 2020.

[205] D. Pfau, J. S. Spencer, A. G. Matthews, and W. M. C. Foulkes, "Ab initio solution of the many-electron schrödinger equation with deep neural networks", *Physical Review Research*, vol. 2, no. 3, p. 033 429, 2020.

[206] L. Richter and J. Berner, "Robust sde-based variational formulations for solving linear pdes via deep learning", in *International Conference on Machine Learning*, PMLR, 2022, pp. 18 649–18 666.

[207] I. Schroeder, "How to resolve microsecond current fluctuations in single ion channels: The power of beta distributions", *Channels*, vol. 9, no. 5, pp. 262–280, 2015.

[208] M. Egerstedt, Y. Wardi, and F. Delmotte, "Optimal control of switching times in switched dynamical systems", in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, IEEE, vol. 3, 2003, pp. 2138–2143.

[209] D. Liberzon, *Switching in systems and control*. Springer Science & Business Media, 2003.

[210] X.-C. Ding, Y. Wardi, and M. Egerstedt, "On-line optimization of switched-mode dynamical systems", *IEEE Transactions on Automatic Control*, vol. 54, no. 9, pp. 2266–2271, 2009.

[211] Y. Wardi, M. Egerstedt, and M. Hale, "Switched-mode systems: Gradient-descent algorithms with armijo step sizes", *Discrete Event Dynamic Systems*, vol. 25, no. 4, pp. 571–599, 2015.

[212] A. Saeedi and A. Bouchard-Côté, "Priors over recurrent continuous time processes", *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[213] F. Stimberg, A. Ruttor, and M. Opper, "Poisson process jumping between an unknown number of rates: Application to neural spike data", *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[214] X. Li, T.-K. L. Wong, R. T. Chen, and D. Duvenaud, "Scalable gradients for stochastic differential equations", in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 3870–3882.

[215] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes", in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2013.

[216] F. Ball, R. Milne, and G. Yeo, "Multivariate semi-Markov analysis of burst properties of multiconductance single ion channels",*Journal of Applied Probability*, vol. 39, no. 1, pp. 179–196, 2002.

[217] N. Engelmann, D. Linzner, and H. Koeppl, "Continuous time bayesian networks with clocks", in *International Conference on Machine Learning*, PMLR, 2020, pp. 2912–2921.

[218] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge university press, 2003.

[219] M. A. Proschan and B. Presnell, "Expect the unexpected from conditional expectation", *The American Statistician*, vol. 52, no. 3, pp. 248–252, 1998.

[220] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, vol. 32, 2019.

[221] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[222] B. Efron, *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012.

ERKLÄRUNG LAUT PROMOTIONSORDNUNG

§ 8 Abs. 1 lit. c PromO
Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§ 8 Abs. 1 lit. d PromO
Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§ 9 Abs. 1 PromO
Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§ 9 Abs. 2 PromO
Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 17. Oktober 2022