WARWICK
THE UNIVERSITY OF WARWICK

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

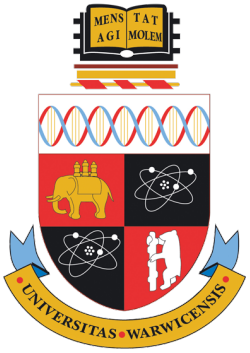
http://wrap.warwick.ac.uk/150539

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

SIMULATION OF CONDITIONED DIFFUSIONS



Marcin Mider

A thesis submitted for the degree of
Doctor of Philosophy

University of Warwick, Department of Statistics

September 2019

Contents

List of Figures

To my parents

Acknowledgements

I would like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC grant number EP/L016710/1) and the University of Warwick for financial support during my studies.

Completing this thesis would have been impossible without the support of my colleagues, peers, friends and family. I would like to thank my supervisors—Paul Jenkins, Murray Pollock and Gareth Roberts—for their guidance and expertise. I was also extremely privileged to have met, collaborated with, learnt from and been inspired by a number of researchers: I would like to thank Michael Sørensen for being my academic host at the University of Copenhagen, for the generosity of his time and for his support; Susanne Ditlevsen for helping me understand many interesting problems in neuroscience; Frank van der Meulen for his numerous insights, comments and discussions on guided proposals; and of course, Moritz Schauer, for his time, conversations, converting me to `Julia` and foremost, showing me how to enjoy academia.

I would like to thank my OxWaSP friends that I started this PhD journey with, for the intense, joint projects that we completed together in our first year in Oxford and for sharing so much knowledge with me.

I have been gifted with the most wonderful and supportive family; had it not been for my parents—Beata and Jarek—the way they raised me, the love and support they extended to me throughout the years, their encouragement and help, I would have never reached this point. Thank you both. And of course, I have to thank my sister—Asia—for her witty remarks and her optimism that spills over to people around her.

Last but not least, my partner—Luisa—thank you for going with me through this journey, for your warmth, your support and your smile that carried me through so many days.

24th September 2019

I would also like to thank my examiners: Wilfrid Kendall and Omiros Papaspiliopoulos for their time, comments, interesting questions and a number of helpful suggestions.

10th January 2020

Declaration

I hereby declare that this thesis is the result of my own work and research, except where otherwise indicated. This thesis has not been submitted for examination to any institution other than the University of Warwick.

Signed: 

Marcin Mider

24th September 2019

Abstract

Diffusion processes are important tools for modelling time-evolution of natural, as well as man-made phenomena. Unfortunately, due to intractability of the likelihood functions performing many standard, statistical analyses for such processes is challenging. For a number of modern, Bayesian techniques the difficulty of dealing with those objects can be reduced to the problem of simulating diffusions conditioned on suitable random variables or events—solving the latter problem is precisely the topic of this thesis. The aim of this thesis is to propose novel extensions to the existing methodologies for simulating diffusions conditioned on an end-point, on partial and noisy observation of the process and on first passage times. I give a comprehensive introduction to this topic in the first part of this document. In the second part I first show how to reduce the computational cost of the so-called *exact* samplers (whose distinguishing feature is lack of any discretisation errors) in the setting of *distant bridges* and I provide a quantitative asymptotic analysis of this cost reduction. Then, I re-formulate the main computational routines of the algorithm termed *guided proposals*, which results in the reduced computational cost of the procedure as well as simplification of its implementation. Finally, I show how to extend a number of algorithms discussed in this thesis to conditioning on first passage times to a threshold. I apply those results to problems encountered in neuroscience: inference from first passage times for leaky integrate-and-fire models as well as for the FitzHugh-Nagumo model.

1 Introduction

Diffusions find a myriad of applications across the engineering, natural and social sciences. They have been used with great success for instance in molecular dynamics to represent proportions of reactants undergoing chemical reactions (Boys et al., 2008) and to model angles between atoms evolving in a force field (Schlick, 2010), in neuroscience to describe the dynamics of the membrane potential of neurons (Lansky and Ditlevsen, 2008) and in astrophysics to approximate quasar variability over time (Kelly et al., 2009). In finance—where they have gained an unparalleled popularity—they are employed to model bond prices, stock prices, options, exchange rates, interest rates and to price a plenitude of other financial instruments (Karatzas and Shreve, 1998b; Di Nunno et al., 2009). In climatology, attempts have been made to use diffusions in describing glacial cycles in energy balance models and variability in the intensity of El Niño and the Southern Oscillation (Imkeller and Monahan, 2002); in biology intracellular processes have been represented with them (Golightly and Wilkinson, 2006); whereas in engineering they foster understanding of the response of various structures to noise, excitation, turbulence and other distortions (Sobczyk, 2013). This is but a small list of problems for which diffusions arise as a natural modelling tool (Kloeden and Platen, 2013; van Zanten, 2013), and in view of the fast-evolving field of numerical techniques for those processes, the list can only expand with time.

Heuristically, diffusions are continuous-time stochastic processes which can be completely characterised by specifying their infinitesimal evolution in time. More formally, a d-dimensional diffusion X is defined as a (Markov) solution to the stochastic differential equation (SDE) of the form:

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \tag{1.1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ denotes the drift coefficient, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d'}$ the volatility coefficient and W a d'-dimensional Brownian motion (known also under the name of the Wiener process), (Øksendal, 2013). Additionally, $\Gamma := \sigma\sigma^T$ is termed the diffusion coefficient. The drift and diffusion coefficients are respectively the infinitesimal mean and covariance of the process X:

$$b(x) = \lim_{\Delta\to 0} \frac{1}{\Delta}\mathbb{E}\big(X_{t+\Delta} - X_t | X_t = x\big),$$

$$\Gamma(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\Big(\big[X_{t+\Delta} - X_t - \Delta b(X_t) \big] \big[X_{t+\Delta} - X_t - \Delta b(X_t) \big]^T \, | X_t = x \Big).$$

Locally, X behaves like a scaled Brownian motion with a drift:

$$X_{t+s} - X_t \approx b(X_t)s + \sigma(X_t)(W_{t+s} - W_t), \quad s \in [0, \Delta], \tag{1.2}$$

though, over longer periods of time the non-linear effects of functions b and σ become pronounced and can give rise to paths strikingly different from those of simple Brownian motion—an example comparing a path of a diffusion based on the Lotka-Volterra equations with a path of Brownian motion is given in fig. 1.1. Notice how the latter wanders about the space aimlessly, in contrast to the former, for which the coordinates of the process regulate each other, keeping the diffusion confined to an ovoid trajectory. Indeed, an important reason contributing to the prevalence of use of diffusions in sciences is that simple manipulations of the drift and volatility coefficients can give rise to an immeasurable diversity of behaviour of the resulting processes.

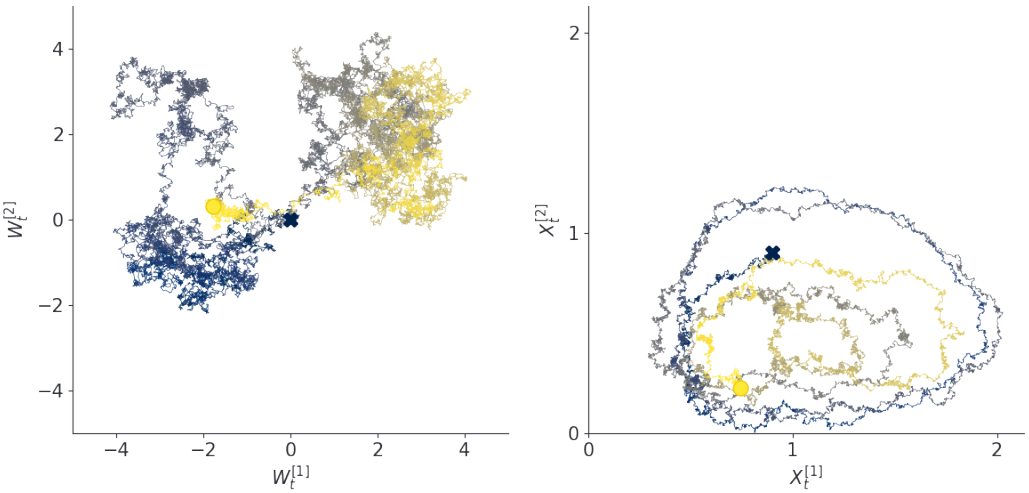


Figure 1.1: A sample path of 2-dimensional Brownian motion $\{W_t, t \in [0, 40]\}$ (left) and a sample path of the stochastic Lotka-Volterra model $\{X_t, t \in [0, 40]\}$ solving eq. (1.1) with $b : \mathbb{R}^2 \to \mathbb{R}^2$, $b^{[1]}(x) := 2x^{[1]}/3 - 4x^{[1]}x^{[2]}/3$, $b^{[2]}(x) := x^{[1]}x^{[2]} - x^{[2]}$ and $\sigma := 0.1 I_2$, where I_2 denotes a 2-dimensional identity matrix and $\cdot^{[i]}$ denotes the i^{th} component of a vector (right). Paths start from the locations marked with crosses and their terminal points are marked with dots. The changing shade illustrates the progression of time.

The primary focus of this thesis is on the problem of simulating conditioned diffusion paths on a computer. More precisely, denoting by $\mathbb{P}$ the law induced by

the SDE (1.1) and by $\mathcal{Z}$ some random variable of interest, the goal is to simulate paths distributed according to $\mathbb{P}(\cdot|\mathcal{Z})$. Some examples of $\mathcal{Z}$ treated in this thesis include:

DIFFUSION BRIDGES $(\mathcal{Z} := X_T)$

Conditioning on an end-point is perhaps the most commonly considered example in the statistical literature. The resulting paths are known under the name of diffusion bridges. Figure 1.2 illustrates this concept for $\mathbb{P}$ being given by the law of Brownian motion.

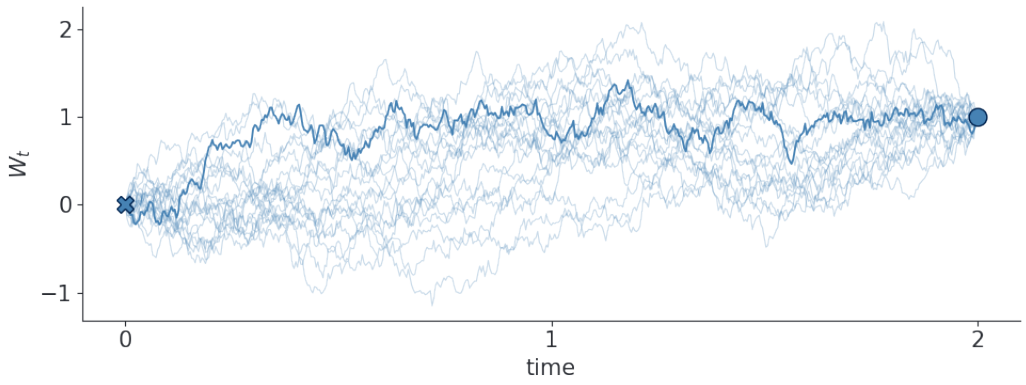


Figure 1.2: Paths of Brownian motion conditioned on an end-point (so-called Brownian bridges). One path is drawn in bold to emphasise a shape of a single trajectory.

PARTIALLY OBSERVED DIFFUSIONS $\left(\mathcal{Z} := \{L_i X_{t_i} + \xi_i,\ i = 1,\ldots,K\}\right)$

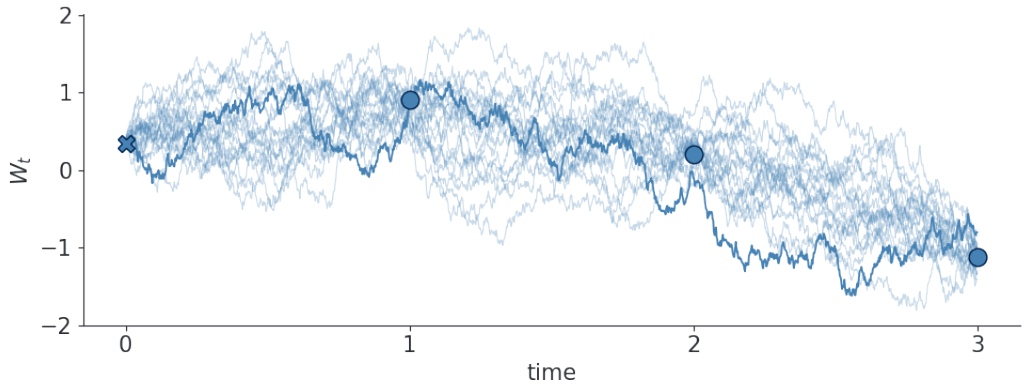


Figure 1.3: Paths of Brownian motion conditioned on three observations $L_i = 1$, $(i = 1,2,3)$ distorted by Gaussian noise: $\xi_i \sim \texttt{Gsn(0,0.2}^2\texttt{)}$, $(i = 1,2,3)$. Observations are given by dots.

with $L_i \in \mathbb{R}^{d_i \times d}$, $d_i \in \mathbb{N}_+$, $\xi_i \sim F_i$ for some known distributions F_i, $(i = 1, \ldots, K)$, $K \in \mathbb{N}_+$. In words, it is conditioning on multiple partial and noisy observations of the process X. Figure 1.3 gives exemplary paths for Brownian motion observed with Gaussian noise.

FIRST-PASSAGE TIME BRIDGES $(\mathcal{Z} := \tau)$

with $\tau(\omega) := \inf\{t \geq 0 : X_t(\omega) \geq L^*\}$; i.e. conditioning on the first passage time to level L^*. Conditioning on the first passage time of Brownian motion is illustrated in fig. 1.4.

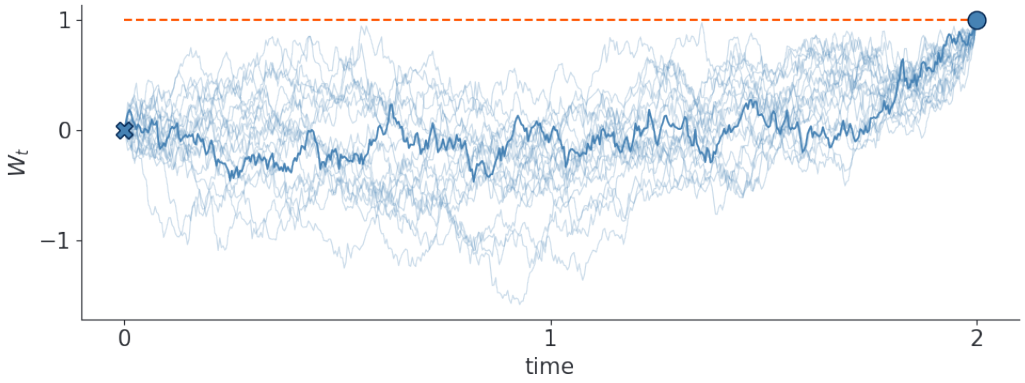


Figure 1.4: Paths of Brownian motion conditioned on the first passage time to level 1.

COMPOSITE FIRST-PASSAGE TIME BRIDGES $(\mathcal{Z} := \tau^*)$

with $\tau_*(\omega) := \inf\{t \geq 0 : X_t^{[1]}(\omega) \leq L_*\}$, $\tau^*(\omega) := \inf\{t \geq \tau_*(\omega) : X_t^{[1]}(\omega) \geq L^*\}$

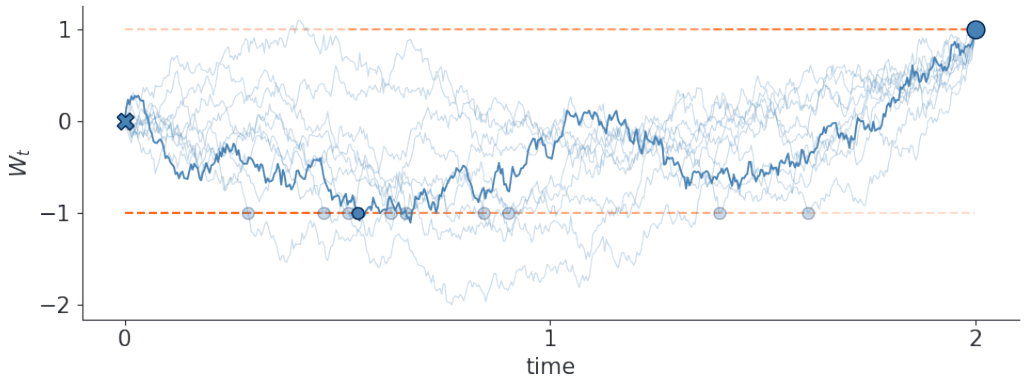


Figure 1.5: Paths of Brownian motion conditioned to be renewed (defined as reaching level −1) at some point during the time interval $[0, 2]$ and conditioned on the first up-crossing of level 1 (post-renewal) to happen at time 2. Notice that before any path is renewed it is allowed to cross an upper threshold 1.

and $L_* < L^*$ (where I henceforth use $^{[i]}$ to refer to an i^{th} coordinate of a vector). This is an elaboration on the first passage time bridges. $\mathcal{Z}$ corresponds to an event of the first coordinate of the process X falling beneath level L_* at some time during $[0, \tau^*]$, staying below L^* from then on, and reaching L^* only at the time τ^*. Figure 1.5 illustrates this conditioning when an underlying process is given by Brownian motion.

Before considering any of those cases however, a more fundamental question needs to be addressed first: since each realisation of the process X is a continuous and intractable function of time—impossible to be stored in its entirety on a computer with finite resources—what is meant exactly by simulating a diffusion path?

The answer is contingent upon the ultimate reason for sampling this object. If the goal is to approximate the values of integrals of the form: $\int_0^T f(X_t)\mathrm{d}t$ or $\int_0^T f(X_t)\mathrm{d}X_t$ or to simply plot realisations of X, it is often enough to fix a dense enough time-grid: $0 = t_0 < t_1 < \cdots < t_M = T$, simulate $(X_{t_0}, X_{t_1}, \ldots, X_{t_M})$ jointly and approximate paths by linear interpolations between X_{t_i}'s. Another objective could be to obtain an unbiased estimate of $\int_0^T f(X_t)\mathrm{d}t$, for which simulation of X_U only at a single time-point $U \in [0, T]$ would suffice[1]. Numerous other use cases exist (some of them will be discussed in this thesis); however, a vast majority of them share a single instrumental component, which leads me to the first definition of this thesis:

Definition 1.0.1. Simulation of a diffusion path is understood as an algorithm capable of jointly sampling $(X_{s_0}, \ldots, X_{s_M})$ for any fixed sequence of time-points $0 \leq s_0 <, \ldots, s_M \leq T$, $M \in \mathbb{N}_+$.

Under certain regularity conditions, a diffusion—which is a strong Markov process—admits the transition density:

$$p_t(x, y)\mathrm{d}y = \mathbb{P}\left(X_{t_0+t} \in \mathrm{d}y | X_{t_0} = x\right).$$

[1]Indeed, $Tf(X_U)$ with $U \sim \texttt{Unif}([0, T])$ could be used as one such unbiased estimator, since:

$$\mathbb{E}[Tf(X_U)] = \mathbb{E}[T\mathbb{E}_{\mathscr{U}}[f(X_U)]] = \mathbb{E}\left[T\int_0^T \frac{1}{T} f(X_t)\mathrm{d}t\right] = \mathbb{E}\left[\int_0^T f(X_t)\mathrm{d}t\right],$$

where $\mathbb{E}_{\mathscr{U}}$ is used to denote the expectation with respect to a uniform random variable.

This provides one idea for a sampling algorithm—the joint density can be decomposed as:

$$p(s_0,\dots,s_M;X_{s_0},\dots,X_{s_M}) = p_{s_0}(X_0,X_{s_0})p_{s_1-s_0}(X_{s_0},X_{s_1})\dots p_{s_M-s_{M-1}}(X_{s_{M-1}},X_{s_M}),$$

and simulation of $(X_{s_0},\dots,X_{s_M})$ proceeds by exploiting the Markov structure of the process: first X_{s_0} is drawn from $p_{s_0}(X_0,X_{s_0})$; then, conditionally on X_{s_0}, X_{s_1} is drawn from $p_{s_1-s_0}(X_{s_0},X_{s_1})$, etc. until all elements $(X_{s_0},\dots,X_{s_M})$ are sampled. Unfortunately, the transition density of virtually all but a handful of the simplest diffusions will be intractable and thus sampling diffusion paths directly from p_t is rarely a viable solution (Kloeden and Platen, 2013, §4.4).

To sample unconditioned diffusions it is instead possible to exploit local behaviour of SDEs. Indeed, for *small* time-step Δ, it follows from eq. (1.2) that $X_{t+\Delta}-X_t$ is approximately distributed as a Gaussian random variable:

$$(X_{t+\Delta}|X_t = x) \sim \texttt{Gsn}(x + b(x)\Delta, \Gamma(x)\Delta),$$

and this approximation gives rise to the celebrated Euler-Maruyama scheme, which I summarise in algorithm 1.1. Unfortunately, the Euler-Maruyama scheme is not suitable for sampling conditioned diffusions which this thesis is concerned with, as the local behaviour of those processes is not only determined by the drift and volatility coefficients, but also by the conditioned-on variable in a way that (in full generality) can no longer be captured by a pair of simple, tractable functions.

Algorithm 1.1 Stochastic Euler-Maruyama scheme

1: Set $s_0 \leftarrow 0$
2: **for** $i = 1,\dots,M$ **do**
3: Set $(x,\Delta) \leftarrow (X_{s_{i-1}}, s_i - s_{i-1})$
4: Draw $X_{s_i} \sim \texttt{Gsn}(x + b(x)\Delta, \Gamma(x)\Delta)$
5: **return** $\{X_{s_i}, i = 1,\dots,M\}$ ▷ Sampled path

Over the years a plethora of alternative methods designed specifically for conditioned diffusions have been proposed in the statistics literature. These methods differ in the assumptions that are imposed on the underlying process X, the computational costs associated with various sampling regimes and the degree of bias that

is introduced into simulations. In this thesis I will describe some of those existing algorithms and build upon them by proposing improvements and introducing their novel extensions.

A substantial part of this thesis is devoted to a discussion of the so-called *exact algorithms*, whose prime characteristic is lack of any discretisation errors. For instance, the Euler-Maruyama scheme would not be classified as an exact algorithm, because it is based on local approximations to SDEs, resulting in a systematic bias in the sampling distribution of $(X_{s_0}, \ldots, X_{s_M})$. On the other hand, direct sampling from tractable transition densities p_t would belong to this category, as no error (beyond the one induced by the floating-point precision of a computer) would be present. Diffusions amenable to exact sampling need to satisfy strict regularity conditions and many computational issues arise when working with those algorithms. In this document I discuss a number of those computational problems and focus in particular on showing how to reduce the computational cost in the setting of simulating long bridges. However, departing from *exactness* property allows one to vastly expand the class of diffusions for which sampling of conditional paths is possible. In the final two chapters I consider a class of algorithms known as guided proposals. This simulation methodology imposes only mild assumptions on the underlying process X, allows for an arbitrary control of induced approximation errors and is robust enough to allow for efficient simulation of conditioned diffusions that lie beyond the capabilities of other, competing methods. My aim is to provide a number of computational improvements to the original formulation of the algorithm and expand it to tackle previously unconsidered conditioned-on variables.

The study of the simulation algorithms for conditioned diffusions is a pivotal component of a number of modern statistical tools designed to work with observational data of phenomena driven by stochastic differential equations. In this thesis I focus primarily on an application to Bayesian inference for diffusion processes. For it, the underlying process is assumed to follow the dynamics of an SDE:

$$\mathrm{d}X_t = b_\theta(X_t)\,\mathrm{d}t + \sigma_\theta(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \tag{1.3}$$

where $\cdot_\theta$ indicates dependence on a vector of unknown and random parameters $\theta \in \Theta$. Parameter vector θ is equipped with a *prior* distribution $\pi(\theta)$ and the process

X is observed at a collection of time-points—I write $\mathscr{D}$ to denote an observation set. Some of the examples of the observation schemes include:

- Exact observations of the process X (i.e. $\mathscr{D} := \{X_{t_i}; i = 1,\ldots,K\}$).
- Partial and noisy observations (i.e. $\mathscr{D} := \{V_{t_i}; i = 1,\ldots,K\}$, with $V_{t_i} = L_i X_{t_i} + \xi_i$, $L_i \in \mathbb{R}^{d_i \times d}$ and $\xi_i \sim F_i$, for some distributions F_i).
- First-passage time observations of a scalar diffusion (i.e. $\mathscr{D} := \{\tau_i; i = 1,\ldots,K\}$, where $\tau_0 := 0$ and $\tau_i := \inf\{t \geq \tau_{i-1} + \epsilon : X_t \geq L^*\}$, $\epsilon > 0$).
- Variations on the first-passage time observations (for instance $\mathscr{D} := \{\tau_i^*; i = 1,\ldots,K\}$, where $\tau_{*0} := 0$, $\tau_i^* := \inf\{t \geq \tau_{*i-1} : X_t^{[1]} \geq L^*\}$, $\tau_{*i} := \inf\{t \geq \tau_i^* : X_t^{[1]} \leq L_*\})$ $i = 1,\ldots,K$, and $L_* < L^*$).
- Exact observation of the entire trajectory: $\mathscr{D} := \{X_t, t \in [0,T]\}$.

The aim is to find the *posterior* density of the unknown parameter vector θ for a given observation set $\mathscr{D}$ and a given prior $\pi(\theta)$:

$$\pi(\theta|\mathscr{D}) \propto \pi(\theta)\pi(\mathscr{D}|\theta).$$

Bayesian algorithms considered in this thesis rely on the step of imputing the unobserved parts of the path, conditionally on a fixed value of the parameter vector θ. To put it differently, it is required to sample paths of diffusions which are solutions to the SDE (1.3) (where θ is just a known vector) and which are conditioned to be consistent with the observation set $\mathscr{D}$. This is precisely what the law $\mathbb{P}(\cdot|\mathscr{Z})$ above denotes, and in fact the entries from the list of examples of $\mathscr{Z}$ are the exact random variables that need to be considered under the corresponding observation regime from the list of examples of $\mathscr{D}$ (with the last example of $\mathscr{D}$ not requiring any path sampling).

1.1 Thesis structure

The thesis is divided into two parts.

In part I, in chapter 2 I provide a succinct summary of the requisite mathematical concepts from Monte Carlo methods and I aim to introduce the corresponding methodologies on a path space as early as possible. This means that even a

reader well-versed in the Monte Carlo methods, but unfamiliar with the methods of sampling on a path space should find chapter 2 insightful. In chapter 3 I give an overview of the current statistical literature relevant for the problem of sampling conditioned diffusions. I single out and describe in detail three methodologies: exact rejection sampling on a path space, simple diffusion bridges and guided proposals. I put the emphasis on the first and the last one of these three because of their importance to further results of this thesis and additionally, because of their unique advantages that make them stand out amongst other diffusion simulation algorithms. I discuss the second method (simple diffusion bridges) so as to posit a correction to an error that is present in the original publication. I also provide a brief (and necessarily, only partial) summary of alternative methods presented in the vast literature on the topic of sampling conditioned diffusions. In chapter 4 I describe a general protocol for Bayesian inference for diffusion processes, based on data augmentation. The centrepiece of this algorithm is the step of simulating conditioned diffusions, so depending on a chosen method for completing this step the overall procedure needs to be appropriately adjusted. Nonetheless, all of the considered inference algorithms share common principles that chapter 4 aims to elicit. Additionally, I provide a brief summary of alternative methods from statistics literature—Bayesian as well as frequentist—used for inference for diffusion processes.

In part II I introduce novel extensions of *exact* algorithms on a path space (where I use the word "exact" to mean "without any discretisation error"), guided proposals as well as other, approximate methods. In chapter 5 I detail the necessary steps for adapting a technique called *blocking*, so as to cast the exact rejection sampler on a path space in a setting of exact Markov chain Monte Carlo algorithm on a path space and as a result, reduce the computational cost of the sampling procedure. I provide a detailed analysis comparing asymptotic computational cost of the two sampling methodologies for simple choices of diffusion processes and illustrate empirically that the conclusions hold in greater generality.

In chapter 6 I reformulate the core computational routines of guided proposals—a versatile methodology that scales well to multidimensional settings and imposes assumptions on the underlying process that are often easier to satisfy in practice—

leading to substantial reductions in the overall computational cost and preparing it for applications to high-dimensional problems. In chapter 7 I consider a problem of inference from first passage time observations. I develop a number of Bayesian inference algorithms (including one exact one) for the leaky integrate-and-fire models with observations of first passage times to some fixed threshold (as is often the case in computational neuroscience). Already at this stage the proposed algorithms go beyond the capabilities of alternative methods of inference suitable for first passage time data that has been described in the statistics literature to this date. By adapting guided proposals I then treat an even more challenging setting of multi-dimensional, hypoelliptic diffusions with observations of first passage times of a single coordinate.

1.2 Thesis contribution

The following is a summary of contributions of this thesis to the statistics literature (in order of appearance in the body of text).

- Quantitative understanding of the cost of blocking (where I define precisely the *blocking* technique in chapter 5).
- A novel computational framework for guided proposals, resulting in substantial reductions in the computational costs.
- A novel, exact Bayesian inference algorithm designed for the leaky integrate-and-fire model with first passage time data.
- Other, novel Bayesian inference algorithms, designed for the leaky integrate-and-fire model with first passage time data.
- A novel Bayesian algorithm applicable to first passage time data of multi-dimensional, hypoelliptic models (applicable in particular to the celebrated FitzHugh-Nagumo model).

PART I

Literature review

2

Preliminaries

2.1 Assumptions

In this thesis I consider various simulation algorithms, which differ in respect of the assumptions that are imposed on the underlying process X. It is instructive to pay close attention to those conditions, as these are often the determining factors for how practically useful a given simulation methodology can be. In this section I will discuss in detail only the most fundamental conditions that most of the algorithms treated in this thesis need to satisfy.

On page 211 I additionally compile a list of all of the assumptions and conditions that appear in the main body of the text. Much of the notation used there is defined only at the point at which a corresponding condition is introduced in the main body of this thesis. The list is supposed to act merely as a reference point, for convenience of the reader.

A solution to a stochastic differential equation (1.1) consists of a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,T]}, \mathbb{P})$, standard, d'-dimensional Brownian motion W and an $\mathscr{F}_t$-adapted, $\mathbb{R}^d$-valued stochastic process X (to which throughout I refer to as a diffusion) with continuous sample paths, such that:

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \quad t \in [0,T],$$

(Le Gall, 2016, Definition 8.1). A solution is weakly unique if for a given triplet X_0, b, σ all solutions have the same law (Karatzas and Shreve, 1998a, §5.3, Definition 3.4). It is strongly unique if for a given quadruplet X_0, b, σ, W and two strong solutions X and $\widetilde{X}$, $\mathbb{P}(X_t = \widetilde{X}_t\,; 0 \le t \le T) = 1$ holds (Karatzas and Shreve, 1998a, §5.2, Definition 2.3). It is non-explosive if $\mathbb{P}(S = \infty) = 1$, where $S := \lim_{n\to\infty} S_n$ and $S_n := \inf\{t \ge 0 : ||X|| > n\}$ (Karatzas and Shreve, 1998a, §5.5, Definition 5.1). One of the most fundamental assumptions required to study diffusions numerically is:

Assumption A1. *Stochastic differential equation* (1.1) *admits a unique, weak, non-explosive solution.*

This is a weak assumption and consequently, a great deal of diffusions of interest will satisfy it. One set of sufficient conditions which implies A1 is (Le Gall, 2016, §8.2):

Condition C1. *Coefficients b and σ are locally Lipschitz continuous and grow at most linearly at infinity. I.e. for any compact set K there exists a constant c_K, such that for all $x, y \in K$:*

$$||\sigma(x)-\sigma(y)|| \leq c_K||x-y||, \quad ||b(x)-b(y)|| \leq c_K||x-y||,$$

and there exists a constant c, such that for all $x, y \in \mathbb{R}^d$:

$$||\sigma(x)|| \leq c(1+||x||), \quad ||b(x)|| \leq c(1+||x||).$$

It is not a necessary condition, though it often suffices for a quick verification. In practice, if condition C1 does not hold it is often because the linear growth condition—which prevents diffusions from exploding to infinity in finite time—is violated.

The double-well potential is an example of a process for which condition C1 does not hold, but which nonetheless satisfies assumption A1:

$$\mathrm{d}X_t = -\rho X_t(X_t^2-\mu)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0,T], \tag{2.1}$$

with $\rho, \sigma > 0$ and $\mu \in \mathbb{R}$ some constants. X experiences a super-linear growth, although, it is only ever directed towards a "correct infinity". As a result, instead of being a force leading to an explosion of X, the cubic term contributes to shrinking of the diffusion's magnitude. Indeed, for $x > 0$: $b(x) := -\rho x(x^2-\mu) < c^*$ for some $c^* \in \mathbb{R}$ and analogously for $x < 0$: $b(x) > c_*$, for some $c_* \in \mathbb{R}$. Consequently, the cubic growth is directed towards the infinity on the other side of the origin and thus acts as a mean-reversion term. The double-well potential is a prototype example for diffusions with coefficients with super-linear growth that nonetheless satisfy A1. A weaker set of conditions that imply A1, designed specifically for such diffusions, can be found for instance in Aït-Sahalia (2002). For yet other tools aiding in verifying existence and uniqueness of solutions to stochastic differential equations see for instance Lipster and Shiryaev (2013, §4.4) or Karatzas and Shreve (1998a, §5).

To perform inference for the class of diffusions with the most general form of the volatility coefficient, a slightly stronger assumption is needed:

Assumption A2. *Stochastic differential equation* (1.1) *admits a unique, strong, non-explosive solution.*

It guarantees existence of a one-to-one mapping between a solution X and a driving Brownian motion W, which turns out to be an essential property requisite for the most general, non-centred implementation of guided proposals (see section 3.3). Thankfully, a wealth of models used in practice satisfy assumption A2—in fact, the same condition C1, which implied the strictly weaker assumption A1, implies A2 as well.

Most algorithms analysed in this thesis are restricted to a uniformly elliptic setting:

Assumption A3. *For all* $x, y \in \mathbb{R}^d$ *there exists an* $\epsilon > 0$ *such that* $y^T \Gamma(x) y \geq \epsilon ||y||^2$.

Although a class of diffusions satisfying this assumption is vast, many examples in natural sciences benefit from lifiting this condition. To this end, I also discuss hypoelliptic diffusions, for which A3 does not hold, but instead A4 does:

Assumption A4. *Solution to stochastic differential equation* (1.1) *admits a smooth density.*

Heuristically, a diffusion is hypoelliptic if its diffusion matrix is not full rank, but is such that its driving noise W affects—at least indirectly—all its coordinates. Formally, A4 can be verified by proving that Hörmander's condition holds (Hörmander, 1967).

Finally, let me remark that in full generality the drift and volatility coefficients may additionally depend on the time variable, so that X could be a solution to the SDE

$$\mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T],$$

instead of (1.1). In fact, an auxiliary diffusion in eq. (3.20), as well as guided proposals in eq. (3.19) are examples of diffusion processes with time-dependent coefficients. To lighten the notation I nonetheless omit explicit dependence on time;

however, the reader should be aware that even in succinct representation of eq. (1.1), the dependence on time is not precluded. Indeed, if the drift and volatility coefficients need to depend on time, one may simply extend the dimension of X by appending the d-dimensional vector X_t with a coordinate $X_t^{[d+1]}$, which solves the degenerate SDE:

$$\mathrm{d}X_t^{[d+1]} = \mathrm{d}t, \quad X_0^{[d+1]} = 0, \quad t \in [0, T].$$

Such extended, $(d+1)$-dimensional process X solves an SDE of the form (1.1) and yet, its coefficients depend on time.[1]

2.2 Girsanov Theorem

The Girsanov theorem is perhaps the single most important result for the field of statistical study of diffusion processes. It gives a recipe for computing the likelihood ratio between two diffusion laws. More precisely, denote by $\mathbb{P}_b$ the law induced by the stochastic differential equation (1.1) and by $\mathbb{P}_\mu$ the law induced by the same SDE but with a different drift $\mu : \mathbb{R}^d \to \mathbb{R}^d$. Suppose that there exists a function $u : \mathbb{R}^d \to \mathbb{R}^{d'}$, such that:

$$\sigma(x)u(x) = b(x) - \mu(x),$$

and that the Novikov's condition A5 holds:

Assumption A5 (Novikov's condition)**.** $\mathbb{E}_\mu\left[\exp\left\{\frac{1}{2}\int_0^T [u^T u](X_t)\,\mathrm{d}t\right\}\right] < \infty$, *where expectation* $\mathbb{E}_\mu$ *is taken with respect to measure* $\mathbb{P}_\mu$.

Then, Girsanov theorem states that the diffusion laws $\mathbb{P}_b$ and $\mathbb{P}_\mu$ are absolutely continuous with respect to one another and that the density of $\mathbb{P}_b$ with respect to $\mathbb{P}_\mu$, evaluated at a path X (drawn under $\mathbb{P}_\mu$) is given by the Radon-Nikodým derivative:

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X) = \exp\left\{\int_0^T u^T(X_s)\,\mathrm{d}W_s - \frac{1}{2}\int_0^T [u^T u](X_s)\,\mathrm{d}s\right\}, \tag{2.2}$$

[1] The coefficients of an SDE are usually allowed to be less regular in the time variable and thus keeping explicit dependence on t is often helpful to derive the most general results.

with W denoting here $\mathbb{P}_\mu$-Brownian motion (Øksendal, 2013, Thmeorem 8.6.5), (Karatzas and Shreve, 1998a, §3.5, Theorem 5.1). This is a powerful statement that makes a notion of the likelihood of a path under a diffusion measure well-defined. However, it turns out that working jointly with the pair (W,X) can be difficult, and instead it is helpful to transform (2.2) to an equivalent statement involving solely path X:

Proposition 2.2.1. (Papaspiliopoulos and Roberts, 2012) If σ is invertible, then (2.2) takes a form:

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X) = \exp\left\{\int_0^T [(b-\mu)^T\Gamma^{-1}](X_s)\,\mathrm{d}X_s - \frac{1}{2}\int_0^T [(b-\mu)^T\Gamma^{-1}(b+\mu)](X_s)\,\mathrm{d}s\right\}. \tag{2.3}$$

Proof. Notice that

$$\begin{aligned} u^T(X_s)\,\mathrm{d}W_s &= [\sigma^{-1}(b-\mu)]^T(X_s)[\sigma^{-1}\sigma](X_s)\,\mathrm{d}W_s \\ &= [(b-\mu)^T(\sigma^{-1})^T\sigma^{-1}](X_s)\sigma(X_s)\,\mathrm{d}W_s \\ &= [(b-\mu)^T(\sigma\sigma^T)^{-1}](X_s)(\mu(X_s)\,\mathrm{d}s + \sigma(X_s)\,\mathrm{d}W_s - \mu(X_s)\,\mathrm{d}s) \\ &= [(b-\mu)^T\Gamma^{-1}](X_s)\,\mathrm{d}X_s - [(b-\mu)^T\Gamma^{-1}\mu](X_s)\,\mathrm{d}s, \end{aligned}$$

and

$$\begin{aligned} [u^T u](X_s)\,\mathrm{d}s &= [(\sigma^{-1}(b-\mu))^T(\sigma^{-1}(b-\mu))](X_s)\,\mathrm{d}s \\ &= [(b-\mu)^T\Gamma^{-1}(b-\mu)](X_s)\,\mathrm{d}s \\ &= [(b-\mu)^T\Gamma^{-1}(b+\mu)](X_s)\,\mathrm{d}s - 2[(b-\mu)^T\Gamma^{-1}\mu](X_s)\,\mathrm{d}s. \end{aligned}$$

Consequently,

$$u^T(X_s)\,\mathrm{d}W_s - \frac{1}{2}[u^T u](X_s)\,\mathrm{d}s = [(b-\mu)^T\Gamma^{-1}](X_s)\,\mathrm{d}X_s - \frac{1}{2}[(b-\mu)^T\Gamma^{-1}(b+\mu)](X_s)\,\mathrm{d}s,$$

and the result follows. □

It should be noted however, that invertibility of σ is not a necessary condition for ridding (2.2) of an explicit dependence on W, as example 2.2.1 below aims to illustrate:

Example 2.2.1. Suppose that $\sigma : \mathbb{R}^2 \to \mathbb{R}^{2\times 1}$ is given by:

$$\sigma(x) := (0, \sigma^*(x))^T,$$

for some positive function $\sigma^* : \mathbb{R}^2 \to \mathbb{R}_+$ (satisfying linear growth condition), and that b and μ are given by:

$$b(x) := (f(x), b^*(x))^T, \quad \mu(x) := (f(x), \mu^*(x))^T,$$

for some sufficiently well-behaved, smooth functions: $f : \mathbb{R}^2 \to \mathbb{R}$, $b^* : \mathbb{R}^2 \to \mathbb{R}$, $\mu^* : \mathbb{R}^2 \to \mathbb{R}$ (so that A1, A4 and A5 hold for $\mathbb{P}_\mu$ and $\mathbb{P}_b$). By Girsanov theorem, (2.2) holds with:

$$u(x) = \left[\frac{b^* - \mu^*}{\sigma^*}\right](x),$$

and although σ is clearly not invertible, it is still possible to remove the explicit dependence on W from (2.2):

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X) = \exp\left\{\int_0^T \left[\frac{b^* - \mu^*}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}X_s^{[2]} - \frac{1}{2}\int_0^T \left[\frac{(b^*)^2 - (\mu^*)^2}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}s\right\}.$$

Proof. To see this, notice:

$$\begin{aligned} u(X_s)dW_s &= \left[\frac{b^* - \mu^*}{\sigma^*}\right](X_s)\left[\frac{\sigma^*}{\sigma^*}\right](X_s)\,\mathrm{d}W_s \\ &= \left[\frac{b^* - \mu^*}{(\sigma^*)^2}\right](X_s)(\mu^*(X_s)\,\mathrm{d}s + \sigma^*(X_s)\,\mathrm{d}W_s - \mu^*(X_s)\,\mathrm{d}s) \\ &= \left[\frac{b^* - \mu^*}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}X_s^{[2]} - \left[\frac{(b^* - \mu^*)\mu^*}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}s, \end{aligned}$$

and

$$\begin{aligned} [u^T u](X_s)\,\mathrm{d}s &= \left[\frac{(b^* - \mu^*)^2}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}s \\ &= \left[\frac{(b^* - \mu^*)(b^* + \mu^*)}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}s - 2\left[\frac{(b^* - \mu^*)\mu^*}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}s. \end{aligned}$$

Consequently,

$$u^T(X_s)\,\mathrm{d}W_s - \frac{1}{2}[u^T u](X_s)\,\mathrm{d}s = \left[\frac{b^* - \mu^*}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}X_s^{[2]} - \frac{1}{2}\left[\frac{(b^*)^2 - (\mu^*)^2}{(\sigma^*)^2}\right](X_s)\,\mathrm{d}s.$$

□

2.3 Rejection sampling

Rejection sampler is an ingenious and very general technique for sampling from one probability measure, using samples from a different probability measure (von Neumann, 1963). More precisely, denote by $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ some density functions, and suppose that there exists a constant $M^* > 0$, s.t.:

$$\frac{f(x)}{M^* g(x)} \leq 1, \quad \forall x \in \mathcal{X}. \tag{2.4}$$

Assume further, that it is possible to obtain iid (independent and identically distributed) draws from g: $X_i \sim g$, $(i = 1, \ldots)$ and that the goal is to sample from f. A rejection sampling protocol takes independent samples from g and accepts them independently with probability given by $f(X_i)/M^* g(X_i)$ (and otherwise rejects them). It turns out that accepted samples are distributed exactly according to f. Algorithm 2.1 below summarises this procedure. To see that the output of

Algorithm 2.1 Rejection sampling

```
while True do
    Draw X ~ g
    Draw U ~ Unif([0, 1])
    if U ≤ f(X)/M* g(X) then
        return X                    ▷ This sample is distributed as X ~ f
```

algorithm 2.1 is indeed a sample from f, define $A := \{X \text{ is accepted}\}$, denote by B any Borel measurable set and notice:

$$\mathbb{P}(X \in A \cap B) = \int_B g(x)\mathbb{P}(X \in A | X = x)\,\mathrm{d}x = \int_B g(x)\frac{f(x)}{M^* g(x)}\,\mathrm{d}x = \frac{\int_B f(x)\,\mathrm{d}x}{M^*}.$$

It follows that under the measure induced by the output of algorithm 2.1, probability of any Borel set B is given by:

$$\mathbb{P}(X \in B | X \text{ accepted}) = \frac{\mathbb{P}(X \in B \cap A)}{\mathbb{P}(X \in A)} = \frac{\int_B f(x)\,\mathrm{d}x}{M^*} \Bigg/ \frac{\int_{\mathcal{X}} f(x)\,\mathrm{d}x}{M^*} = \int_B f(x)\,\mathrm{d}x,$$

showing that $(X|X \text{ accepted})$ is indeed a sample from f.

2.3.1 Overview of rejection sampling on a path space

Rejection sampling on a path space is simply a special case of the rejection sampler above (Beskos and Roberts, 2005), in which $\mathscr{X} = \mathcal{C}([0,T];\mathbb{R}^d)$ is a space of diffusion paths ($\mathcal{C}(A;B)$ denotes a space of continuous functions form $A \to B$) and f and g are two conditioned diffusion measures $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$ and $\mathrm{d}\mathbb{P}_\mu(\cdot|\mathcal{Z})$ (see Introduction for some examples of the conditioned-on random variables $\mathcal{Z}$). The ratio of the two densities f and g is now given by the Radon-Nikodým derivative between the two measures: $f(x)/g(x) = [\,\mathrm{d}\mathbb{P}_b/\mathrm{d}\mathbb{P}_\mu](x|\mathcal{Z})$ and M^* is a global upper bound on this ratio:

$$\frac{1}{M^*}\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X|\mathcal{Z}) \leq 1, \quad \forall X \in \mathcal{C}([0,T];\mathbb{R}^d), \text{ which are consistent with } \mathcal{Z}. \tag{2.5}$$

The algorithm proceeds by drawing paths from $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ and accepting them with probability (2.5). This is summarised in algorithm 2.2 below.

Algorithm 2.2 Rejection sampling on a path space

1: **while** `True` **do**
2: Draw $X \sim \mathrm{d}\mathbb{P}_\mu(\cdot|\mathcal{Z})$
3: Draw $U \sim \mathtt{Unif}([0,1])$
4: **if** $M^*U \leq [\,\mathrm{d}\mathbb{P}_b/\mathrm{d}\mathbb{P}_\mu](X|\mathcal{Z})$ **then**
5: **return** X ▷ This sample is distributed as $X \sim \mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

2.3.2 Lamperti transformation & potential form

To reiterate, algorithm 2.2 makes it possible to draw paths under some conditioned diffusion measure $\mathbb{P}_b(\cdot|\mathcal{Z})$, so long as:

Condition C2. *It is possible to obtain draws from* $\mathbb{P}_\mu(\cdot|\mathcal{Z})$.

Condition C3. *There exists a constant* M^* *satisfying* (2.5).

Condition C4. *The Radon-Nikodým derivative* $\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X|\mathcal{Z})$ *exists and is tractable.*

This means that sometimes the problem of drawing from $\mathbb{P}_b(\cdot|\mathcal{Z})$ for a given, fixed b can be reduced to a simpler problem of drawing from $\mathbb{P}_\mu(\cdot|\mathcal{Z})$, where μ is a free function-parameter that can be chosen by a practitioner. Unfortunately, as it

stands, finding a global upper bound M^* seems impossible. Recall expressions (2.2) and (2.3) for the Radon-Nikodým derivative between two unconditioned diffusion laws. As I will show in eq. (2.10), this term appears in a product form for an expression for $[\mathrm{d}\mathbb{P}_b/\mathrm{d}\mathbb{P}_\mu](X|\mathcal{Z})$ and contains an Itō integral, which is unbounded on $X \in \mathcal{C}([0,T];\mathbb{R}^d)$. To this end a sequence of transformations—starting with Lamperti transformation (Kloeden and Platen, 2013, §4.4)—needs to be employed.

Assumption A6 (Lamperti transformation)**.** *σ is invertible and there exists a function $\eta : \mathbb{R}^d \to \mathbb{R}^d$, such that: $(\nabla_x \eta(x))^T = \sigma^{-1}(x)$.*

Suppose that assumption A6 holds. The Lamperti transformed diffusion is defined as $Y := \{\eta(X_t); t \in [0,T]\}$, and it solves the following stochastic differential equation:

$$\mathrm{d}Y_t = \alpha(Y_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad Y_0 = y_0, \quad t \in [0,T], \tag{2.6}$$

where $y_0 := \eta(x_0)$, $\alpha^{[k]}(y) := \mathscr{L}\eta^{[k]}(\eta^{-1}(y))$, $(k = 1,\ldots,d)$ and

$$\mathscr{L}f(x) := [b(x)]^T \nabla_x f(x) + \frac{1}{2}\Gamma(x) : \nabla_x \nabla_x f(x), \tag{2.7}$$

is the infinitesimal generator of the process X (where $\Gamma := \sigma\sigma^T$ and : denotes the Frobenius inner product[2]). Many examples in this thesis involve scalar diffusions (for them A6 is always satisfied), so it is instructive to keep in mind the form of the new drift α in this particular setting:

$$\alpha(y) := \frac{b}{\sigma}\left(\eta^{-1}(y)\right) - \frac{1}{2}\sigma'\left(\eta^{-1}(y)\right).$$

Notice that a sample path X can be obtained from a sample path Y by virtue of inverting the Lamperti transformation $X := \{\eta^{-1}(Y_t); t \in [0,T]\}$. Consequently, so long as A6 holds, SDE (1.1) inducing law $\mathbb{P}_b$ can be assumed to have the volatility coefficient equal to an identity matrix: $\sigma = I_d$. This in turn means that if A6 and C2–C4 hold, in order to sample paths of a general diffusion with some pre-defined drift and volatility coefficients, conditionally on the variable $\mathcal{Z}$, it is enough to be able to sample from the conditioned diffusion law $\mathbb{P}_\mu(\cdot|\mathcal{Z})$, where μ is a free function-parameter and the volatility coefficient is just an identity matrix.

[2]Frobenius inner product between two real valued matrices A and B is given by: $A : B := \sum_{ij} A_{ij}B_{ij} = \mathrm{tr}(A^T B)$. Notice that if $B = vv^T$ for some vector $v \in \mathbb{R}^d$, then $A : B = \mathrm{tr}(A^T vv^T) = \mathrm{tr}(v^T A^T v) = v^T A^T v$.

Given tractability of Brownian motion's transition densities and an ample literature about the processes related to Brownian motion through various conditionings (see Borodin and Salminen (2002)), a choice $\mu = 0$ is often made in practice.[3] Suppose further that A7 and A8 hold:

Assumption A7 (Potential form)**.** *The drift of a Lamperti transformed diffusion is of the potential form: i.e. there exists (a potential function)* $A : \mathbb{R}^d \to \mathbb{R}$*, such that* $\nabla_x A(x) = \alpha(x)$*.*

Assumption A8. *The drift of a Lamperti transformed diffusion is continuously differentiable:* $\alpha \in \mathcal{C}^1(\mathbb{R}^d; [0, T])$*.*

Then, the Radon-Nikodým derivative between the laws $\mathbb{P}_\alpha$ and $\mathbb{P}_0$ (induced by (2.6) and d-dimensional Brownian motion respectively) is given by:

$$\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y) = \exp\left\{A(Y_T) - A(Y_0) - \int_0^T \varphi(Y_s)\,\mathrm{d}s\right\}, \tag{2.8}$$

$$\text{where} \quad \varphi(x) := \frac{1}{2}\left(||\alpha(x)||^2 + \Delta_x A(x)\right),$$

(Dacunha-Castelle and Florens-Zmirou, 1986). In dimension 1, φ function takes the following form:

$$\varphi(x) := \frac{1}{2}\left(\alpha^2(x) + \alpha'(x)\right). \tag{2.9}$$

Proof. Starting from (2.3) and using stochastic integration by parts:

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y) &= \exp\left\{\int_0^T \alpha^T(Y_s)\,\mathrm{d}Y_s - \frac{1}{2}\int_0^T [\alpha^T\alpha](X_s)\,\mathrm{d}s\right\} \\ &= \exp\left\{A(Y_T) - A(Y_0) - \frac{1}{2}\int_0^T \Delta_y A(Y_s)\,\mathrm{d}s - \frac{1}{2}\int_0^T ||\alpha(Y_s)||^2\,\mathrm{d}s\right\}.\end{aligned}$$

□

As the stochastic integral has been successfully eliminated, it is reasonable to expect that under certain regularity conditions on α, the ratio of densities $[\mathrm{d}\mathbb{P}_\alpha/\mathrm{d}\mathbb{P}_0](Y)$ is bounded uniformly in $Y \in \mathcal{C}([0, T]; \mathbb{R}^d)$. It is left to connect this ratio with its conditioned counterpart: $[\mathrm{d}\mathbb{P}_\alpha/\mathrm{d}\mathbb{P}_0](Y|\mathcal{Z})$.

[3] Wright-Fisher diffusion is an exception to this rule—to sample the Wright-Fisher diffusion with some general drift b via rejection sampling on a path space, the proposal process is also taken to be the Wright-Fisher diffusion, though with a drift μ chosen carefully enough so that sampling $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ directly from the transition densities is possible, see Jenkins and Spano (2017).

2.3.3 Factorisation of measures & path space rejection sampling

Consider the following mild abstraction of the argument of Roberts and Stramer (2001). Assume that A9 below holds.

Assumption A9. *$\mathcal{Z}$ admits densities under the proposal $\mathbb{P}_\mu$ and the target $\mathbb{P}_b$ laws with respect to some dominating measure $\varrho(\mathrm{d}\mathbb{z})$.*

Write $g_\mu(\mathbb{z})$ and $g_b(\mathbb{z})$ respectively to denote these two densities:

$$g_\mu(\mathbb{z})\varrho(\mathrm{d}\mathbb{z}) := \mathbb{P}_\mu(\mathcal{Z} \in \mathrm{d}\mathbb{z}), \quad g_b(\mathcal{Z})\varrho(\mathrm{d}\mathbb{z}) := \mathbb{P}_b(\mathcal{Z} \in \mathrm{d}\mathbb{z}).$$

Then, the proposal and the target laws can be factorised as follows:

$$\mathbb{P}_\mu = \mathbb{P}_\mu(\cdot|\mathcal{Z} = \mathbb{z}) \otimes g_\mu(\mathbb{z})\varrho(\mathrm{d}\mathbb{z}), \quad \mathbb{P}_b = \mathbb{P}_b(\cdot|\mathcal{Z} = \mathbb{z}) \otimes g_b(\mathbb{z})\varrho(\mathrm{d}\mathbb{z}),$$

where I use $\otimes$ to define product measures. As a result:

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X|\mathcal{Z}) = \frac{g_\mu(\mathcal{Z})}{g_b(\mathcal{Z})}\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X) \propto \frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X). \tag{2.10}$$

This makes it possible to formulate rejection sampling on a path space more explicitly.

Theorem 2.3.1. Suppose that A1, A3 and A5–A9 hold and that α is sufficiently regular so that eq. (2.5) holds. Then, the rejection sampling algorithm 2.3 below outputs paths distributed according to $\mathbb{P}_b(\cdot|\mathcal{Z})$:

Algorithm 2.3 Rejection sampling on a path space via conditioned Brownian motion

while `True` **do**
  Draw $Y \sim \mathrm{d}\mathbb{P}_\alpha(\cdot|\mathcal{Z})$ ▷ Brownian motion conditioned on $\mathcal{Z}$
  Draw $E \sim \mathtt{Exp}(1)$
  if $E \geq -A(Y_T) + A(Y_0) + \int_0^T \varphi(Y_s)\,\mathrm{d}s + \log M^*$ **then**
    Set $X \leftarrow \{\eta^{-1}(Y_t); t \in [0, T]\}$
    return X ▷ This sample is distributed as $X \sim \mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

Proof. The theorem follows from the discussion above and a well known simulation result, stating that $U \sim \mathtt{Unif}([0,1])$ if an only if $-\log(U) \sim \mathtt{Exp}(1)$ (Devroye, 2006). □

Example 2.3.1. (Beskos and Roberts, 2005) Suppose that $\mathcal{Z} := X_T$, i.e. that the goal is to sample a diffusion bridge (for a diffusion with the drift b and the volatility coefficient σ) and that A1, A3 and A5–A8 hold (assumption A9 holds by virtue of existence of transition densities). Notice that conditionally on $\mathcal{Z}$, the end-point $Y_T := \eta(X_T)$ is just a constant and hence (2.10) takes the following form:

$$\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y|\mathcal{Z}) \propto \exp\left\{-\int_0^T \varphi(Y_s)\,\mathrm{d}s\right\}.$$

C3 asserting existence of a global upper bound M^* is now equivalent to the following assumption A10:

Assumption A10. *There exists a constant $l_* > \infty$, s.t. $l_* \leq \inf\{\varphi(x);\, x \in \mathbb{R}^d\}$.*

Consequently, acceptance probability takes the form:

$$\frac{1}{M^*}\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y|\mathcal{Z}) = \exp\left\{-\int_0^T \phi(Y_s)\,\mathrm{d}s\right\}, \tag{2.11}$$

where:

$$\begin{aligned} \phi(x) &:= \varphi(x) - l_* = \frac{1}{2}\left(||\alpha(x)||^2 + \Delta_x A(x)\right) - l_*, \quad \text{and} \\ M^* &:= \frac{g_\mu(\mathcal{Z})}{g_b(\mathcal{Z})}\exp(A(Y_T) - A(Y_0) - l_* T). \end{aligned} \tag{2.12}$$

Now, algorithm 2.3 may be fully applied to sample from $\mathbb{P}_b(\cdot|\mathcal{Z})$. Line 4 of algorithm 2.3 prompts for evaluation of $\int_0^T \phi(Y_s)\,\mathrm{d}s$. On the first attempt, for any realisation of a path Y, I can simply approximate the value of such integral with Riemann sums: $\int_0^T \phi(Y_{s_i})\,\mathrm{d}s \approx \sum_{i=0}^N \phi(Y_s)\frac{s_{i+1}-s_i}{T}$, where $0 = s_0 < \cdots < s_M < s_{M+1} = T$ is a sufficiently dense grid. Therefore, in line 2 of algorithm 2.3 (which for this choice of $\mathcal{Z}$ boils down to sampling of Brownian bridges), it is enough to reveal path Y merely at the grid locations $s_0, \ldots, s_M$. A well-known sampler can be employed for this step. It is based on the following identity (Karatzas and Shreve, 1998a, §5.5.6, Problem 6.14)

$$B \overset{d}{=} \left\{x_0 + W_t + (x_T - x_0 - W_T)\frac{t}{T};\, t \in [0,T]\right\}, \tag{2.13}$$

where B denotes a Brownian bridge joining x_0 and x_T and W is an independent, d-dimensional Brownian motion. Algorithm 2.4 summarises this last sampler.

Algorithm 2.4 Sampler of Brownian bridges

1: Draw a Brownian motion $(W_{s_0}, \ldots, W_{s_M}, W_T)$
2: Set $(B_{s_0}, \ldots, B_{s_M}) \leftarrow x_0 + (W_{s_0}, \ldots, W_{s_M}) + \frac{x_T - x_0 - W_T}{T}(s_0, \ldots, s_M)$
3: **return** $(B_{s_0}, \ldots, B_{s_M})$ ▷ Brownian bridge

Surprisingly, it is possible to do better than that. An approximation via Riemann sums is a weak link of the discussed implementation of algorithm 2.3, and it introduces an approximation error. However, it turns out to be possible to implement algorithm 2.3 in such a way that no approximation ever needs to be made. This is the subject of the exact algorithm of Beskos and Roberts (2005); Beskos et al. (2006, 2008) discussed in section 3.1.

Example 2.3.2. (Beskos and Roberts, 2005) Suppose that $\mathcal{Z} := \varnothing$, i.e. that the goal is to sample an unconditioned diffusion with the drift b and the volatility σ and that A1, A3, A5–A8 and A10 hold. In view of the very efficient samplers based on stochastic Taylor expansions (Kloeden and Platen, 2013)—of which the Euler-Maruyama scheme given in algorithm 1.1 is an example—it might seem wasteful to consider $\mathcal{Z} := \varnothing$; however, much like example 2.3.1 is a prelude to an algorithm capable of exact sampling of diffusion bridges, so is this one for exact sampling of unconditioned diffusions.

In this example $Y_T := \eta(X_T)$ is no longer a constant and this makes proposals from $\mathbb{P}_0$ sub-optimal. Instead, start from assuming:

Assumption A11. *Function* $\upsilon(y) := \exp\{-||y-z||^2/2T + A(y)\}$ *is integrable for any* $z \in \mathbb{R}^d$.

And define the measure of a *biased Brownian motion* $\mathbb{Z}$, through the Radon-Nikodým derivative with respect to $\mathbb{P}_0$:

$$\frac{\mathrm{d}\mathbb{Z}}{\mathrm{d}\mathbb{P}_0}(Y) \propto \exp\{A(Y_T)\}.$$

The naming convention is clear after realising that conditionally on an end-point the two laws coincide: $\mathbb{Z}(\cdot|Y_T) = \mathbb{P}_0(\cdot|Y_T)$, and that under $\mathbb{Z}$, an end-point is distributed according to υ:

$$\mathbb{Z}(Y_T \in \mathrm{d}y) = \upsilon(y)\,\mathrm{d}y.$$

Consequently, if $\mathbb{Z}$ is used as a proposal law, then acceptance probability is of the same form as in example 2.3.1:

$$\frac{\mathrm{d}\mathbb{P}_{\alpha}}{\mathrm{d}\mathbb{Z}}(Y) = \frac{\mathrm{d}\mathbb{P}_{0}}{\mathrm{d}\mathbb{Z}}(Y)\frac{\mathrm{d}\mathbb{P}_{\alpha}}{\mathrm{d}\mathbb{P}_{0}}(Y) \propto \exp\left\{-\int_{0}^{T}\phi(Y_s)\,\mathrm{d}s\right\}. \tag{2.14}$$

The simulation algorithm takes the following form:

Algorithm 2.5 Rejection sampling on a path space for unconditioned diffusions

1: **while** `True` **do**
2: Draw $Y_T \sim \upsilon$
3: Draw $Y \sim \mathbb{P}_0(\cdot|Y_T)$ ▷ Law of a Brownian bridge, use algorithm 2.4
4: Draw $E \sim \mathtt{Exp}(1)$
5: **if** $E \geq \int_0^T \phi(Y_s)\,\mathrm{d}s$ **then**
6: Set $X \leftarrow \{\eta^{-1}(Y_t); t \in [0, T]\}$
7: **return** X ▷ This sample is distributed as $X \sim \mathrm{d}\mathbb{P}_b$

Akin to example 2.3.1, Riemann sums can be used to approximate integrals $\int_0^T \phi(Y_s)\,\mathrm{d}s$. This prompts for revealing path Y in line 3 of algorithm 2.5 only at a dense enough time-grid: $(s_0, \ldots, s_M)$. Naturally, algorithm 2.5 with Riemann sum approximation does not improve upon competing algorithms based on stochastic Taylor expansions. However, in section 3.1 I demonstrate a modification due to Beskos and Roberts (2005); Beskos et al. (2006, 2008), which removes all approximation errors and constitutes an improvement over the other samplers of unconditioned diffusions.

2.3.4 Issues with rejection sampling on a path space

Performance of any rejection sampling algorithm is contingent upon the fidelity with which a proposal law approximates the target. Figure 2.1 illustrates this concept. If a proposal density g puts a great deal of mass on regions of space with low probability mass under f, then most samples from g are improbable under f (this manifests itself in very small values of acceptance probability (2.4) and excessively high rejection rates). Conversely, if the two densities f and g match each other closely on an entire space, then (2.4) remains close to 1, leading to low rejection rates and an efficient sampler.

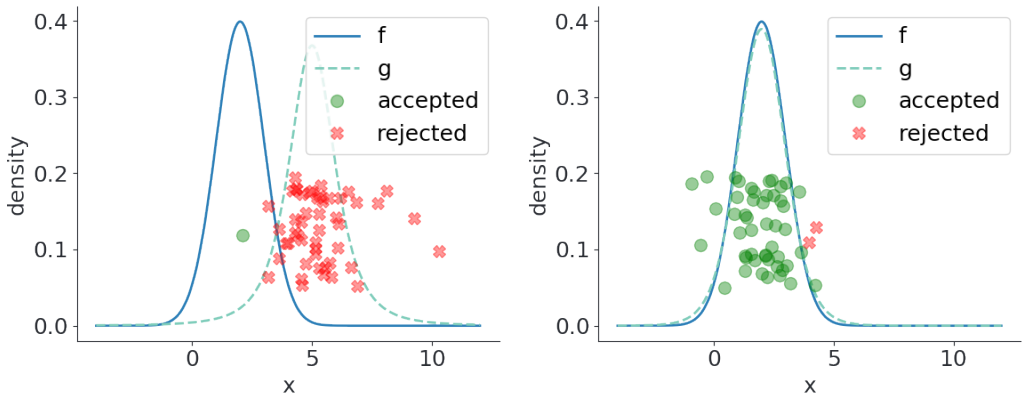


Figure 2.1: Illustration of an impact that discrepancies between proposal and target laws have on the performance of rejection sampling. In both plots target f is set to $\texttt{Gsn}(2, 1^2)$. On the left, proposal g is set to $\texttt{Student-t}(3)$ shifted to the right by 5. On the right, proposal g is set to $\texttt{Student-t}(10)$ shifted to the right by 2. Density plots are given, together with 50 random samples from g (scattered randomly in the y-plane for better transparency). Samples accepted during a test run of a rejection sampler targeting f are marked with green dots, rejected samples are marked with red crosses. Rejection sampler on the right, where f and g are closely matched, accepts $\sim 94\%$ of the proposed samples, clearly outperforming the setting on the left, where most samples ($\sim 96\%$) are rejected.

In the context of rejection sampling on a path space this means that the proposal law $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ needs to be *close* to the target $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$. Unfortunately, as argued in section 2.3.2 or example 2.3.2 the choice of $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ is often heavily restricted due to computational considerations, and this means that finding a better proposal is rarely a viable option for improving performance of the algorithm. Consequently, the more $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$ departs from the law of Brownian motion, the more inefficient rejection sampling on a path space becomes.

A profound implication of this observation follows from recalling eq. (1.2), which states that locally, solution to any SDE behaves as a scaled Brownian motion with a drift. Heuristically, it means that for a *small* T the two laws $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$ and $\mathbb{P}_0(\cdot|\mathcal{Z})$ are close, whereas as $T \to \infty$, they become increasingly dissimilar. In section 3.1.4 I show that the discrepancy increases exponentially quickly in T, leading to an exponential explosion in the cost of the sampler.

For unconditioned diffusions this observation can be used to one's advantage through a simple application of the Markov property. The interval $[0, T]$ is chopped into smaller sub-intervals: $0 = t_0 < \cdots < t_M < t_{M+1} = T$ and sampling procedure

is broken down into a sequence of steps. First, a segment of Y on $[t_0, t_1]$ is drawn from $\mathbb{P}_0|_{t_1}$ to target $\mathbb{P}_\alpha|_{t_1}$ ($\mathbb{P}_0|_{t_1}$ denotes a restriction of $\mathbb{P}_0$ to $\mathscr{F}_{t_1}$). Upon acceptance, the next segment on $[t_1, t_2]$ is drawn from $\mathbb{P}_0|_{t_2}(\cdot|\mathscr{F}_{t_1})$ (a restriction of $\mathbb{P}_0$ to $\mathscr{F}_{t_2}$ conditioned on $\mathscr{F}_{t_1}$—i.e. conditioned on a path simulated until time t_1) and targets $\mathbb{P}_\alpha|_{t_2}(\cdot|\mathscr{F}_{t_1})$. If the path is rejected, then only a new proposal from $\mathbb{P}_0|_{t_2}(\cdot|\mathscr{F}_{t_1})$ needs to be re-drawn, and Y on $[t_0, t_1]$ remains unaffected. This continues until Y is sampled on all sub-intervals. By the Markov property, the final path Y is distributed according to $\mathbb{P}_\alpha$. This brings down the computational cost of path space rejection sampler for unconditioned diffusions down to linear in T.

Unfortunately, the same technique cannot be used in general, as the conditioned-on random variable $\mathcal{Z}$ might depend on the parts of the path outside of $\mathscr{F}_{t_i}$. Indeed, take diffusion bridges or first passage time bridges (see Introduction for definitions) as examples. In both cases sampling Y on $[t_0, t_1]$ requires computation of $\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}|_{t_1}(Y|\mathcal{Z})$. Since $\mathcal{Z}$ depends on Y_T this step is intractable.

A substantial part of this thesis is devoted to addressing this computational hurdle. In chapter 5 I show that it is possible to modify the path space rejection sampler by employing the *blocking* technique, bringing down the scaling of the sampler's computational cost down to cubic in T. Nonetheless, this comes at a cost of having to define an algorithm in the context of Markov chain Monte Carlo sampler, presented in section 2.5 below.

Rejection sampling on a path space has a second drawback still and it has to do with the imposed assumptions: both A6 and A7 are restrictive. Although in dimension 1 the two are satisfied so long as A3 holds, this changes in a multivariate setting. Indeed, consider the following example:

Example 2.3.3. Suppose that:

$$\alpha(x) := \begin{pmatrix} c_1 x^{[2]} + b_1^*(x) \\ c_2 x^{[1]} + b_2^*(x) \\ b_3^*(x) \end{pmatrix},$$

where $c_1, c_2 \in \mathbb{R}$ are some constants, $b_i^* : \mathbb{R}^d \to \mathbb{R}$, $(i = 1, 2)$ are some functions which do not include any linear terms in $x^{[i \bmod 2+1]}$ $(i = 1, 2)$, and $b_3^* : \mathbb{R}^d \to \mathbb{R}^{d-2}$. If a potential A of α exists, then in particular it must hold:

$$\frac{\partial A}{\partial x^{[1]}}(x) = c_1 x^{[2]} + b_1^*(x), \quad \frac{\partial A}{\partial x^{[2]}}(x) = c_2 x^{[1]} + b_2^*(x),$$

and thus:

$$c_1 x^{[1]} x^{[2]} + \int_0^{x^{[1]}} b_1^*\left(\begin{pmatrix} u^{[1]} \\ x^{[-1]} \end{pmatrix}\right) \mathrm{d}u^{[1]} + c_1^*(x^{[-1]}) = c_2 x^{[1]} x^{[2]} + \int_0^{x^{[2]}} b_2^*\left(\begin{pmatrix} x^{[1]} \\ u^{[2]} \\ x^{[3:d]} \end{pmatrix}\right) \mathrm{d}u^{[2]} + c_1^*(x^{[-2]}), \tag{2.15}$$

where $v^{[-j]} := (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)$, $v^{[j,k]} := (v_j, \ldots, v_k)$ and $c_i^* : \mathbb{R}^{d-1} \to \mathbb{R}$, $i = 1, 2$ are some functions. The above can be true only if $c_1 = c_2$. What it means for rejection sampling of multivariate diffusions is that if the dynamics of one coordinate are influenced by another coordinate in a linear manner, then the same, symmetric influence must be exerted by the first coordinate on the latter. This is an undesirable restriction that most multivariate models used in practice violate (an exception is the multivariate double-well potential, or more generally, examples from physics with the distinguishing features that the energy of the system is conserved). The reason why multivariate models are used in practice is to couple the dynamics of the coordinates so as to create systems with behaviours that are impossible to re-create in a 1-dimensional setting. For a handsome number of diffusion models the coupling of dynamics is done exclusively through linear terms of other coordinates, without needing to include non-linear or interaction terms. However, the imposed symmetry $c_1 = c_2$ is rarely satisfied. Naturally, any hope of existence of a potential function A in the even more complicated models, in which interdependence of coordinates goes beyond linear, decreases further yet.

A very similar argument to the one in example 2.3.3 can be made for assumption A6; however, since a plethora of interesting models exist for which the diffusion coefficient σ has a simple form, focusing on relaxing A6 is a secondary issue. With that being said, a simple diffusion that is defined on a complicated manifold will inevitably violate A6, as the local curvature of a manifold is expressed through the modified coefficients of an SDE (Rogers and Williams, 2000b, Ch.V). Consequently, rejection sampling on a path space is not immediately applicable to such models.

2.4 Importance sampling

Heuristically, importance sampler is a rejection sampler for which rejection step is substituted with weighting. More precisely, suppose as before that $f : \mathscr{X} \to \mathbb{R}$ and $g : \mathscr{X} \to \mathbb{R}$ are some density functions, and that the following absolute continuity result holds: for any Borel set B if $\int_B g(x)\mathrm{d}x = 0$, then $\int_B f(x)\mathrm{d}x = 0$ (written $f \ll g$). Assume further that it is possible to obtain iid draws from g: $X_i \sim g$, $(i = 1, \ldots)$ and that functions f and g can be evaluated only up to normalisation constants, i.e. only $f_u : \mathscr{X} \to \mathbb{R}$ and $g_u : \mathscr{X} \to \mathbb{R}$ are known, with $f(x) = c_1 f_u(x)$, $g(x) = c_2 g_u(x)$ and $c_1, c_2 > 0$ some unknown constants. The goal is to compute integrals of the form:

$$f(\pi) := \int_{\mathscr{X}} \pi(x) f(x)\mathrm{d}x,$$

for some f-integrable functions $\pi : \mathscr{X} \to \mathbb{R}$. Importance sampling is based on the following observation:

$$f(\pi) := \int_{\mathscr{X}} \pi(x) f(x)\mathrm{d}x = \int_{\mathscr{X}} \pi(x) \frac{c_1 f_u}{c_2 g_u}(x) g(x)\mathrm{d}x = \frac{c_1}{c_2} g(\pi w),$$

where $w(x) := \frac{f_u}{g_u}(x)$ is a weighting function (Kahn and Harris, 1951). Since

$$g(w) := \int_{\mathscr{X}} w(x) g(x)\mathrm{d}x = \int_{\mathscr{X}} \frac{c_2}{c_1} \frac{f(x)}{g(x)} g(x)\mathrm{d}x = \frac{c_2}{c_1} \int_{\mathscr{X}} f(x)\mathrm{d}x = \frac{c_2}{c_1},$$

it follows that

$$f(\pi) = \begin{cases} g(\pi w) & \text{if } c_1 = c_2, \\ \frac{g(\pi w)}{g(w)} & \text{else.} \end{cases}$$

Therefore, integrating function π with respect to density f is equivalent to integrating function πw with respect to another density g and scaling the result by $1/g(w)$. If $f(|w\pi|) < \infty$, then, by Slutsky's lemma and the strong law of large numbers:

$$\begin{cases} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N \pi(X_i) w(X_i) = f(\pi) \quad \text{(a.s.)}, \quad \text{for } X_i \sim g, & \text{if } c_1 = c_2, \\ \lim_{N\to\infty} [\sum_{j=1}^N w(X_j)]^{-1} \sum_{i=1}^N \pi(X_i) w(X_i) = f(\pi) \quad \text{(a.s.)}, \quad \text{for } X_i \sim g, & \text{else .} \end{cases}$$

This is the basis for defining the following Monte Carlo estimators:

$$f(\pi) \approx \begin{cases} \frac{1}{N} \sum_{i=1}^N \pi(X_i) w(X_i), \\ [\sum_{j=1}^N w(X_j)]^{-1} \sum_{i=1}^N \pi(X_i) w(X_i), \end{cases}$$

where $X_i \sim g$, $(i = 1, \ldots, N)$. The first estimator is unbiased:

$$\mathbb{E}_g\left[\frac{1}{N}\sum_{i=1}^{N}\pi(X_i)w(X_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_g\left[\pi(X_1)w(X_1)\right] = \mathbb{E}_f\left[\pi(X_1)\right],$$

however, it is the second—the so-called self-normalised estimator—that is more useful in practice, as normalisation constants of f and g do not need to be known. Algorithm 2.6 below summarises an importance sampling procedure. Weighted particles returned as the output of algorithm 2.6 can then be used to estimate $f(\pi)$.

Algorithm 2.6 Importance sampling

1: **for** $i = 1, \ldots, N$ **do**
2: Draw $X_i \sim g$
3: Set $w_i \leftarrow \frac{f_u}{g_u}(X_i)$
4: **return** $\{(X_i, w_i); i = 1, \ldots, N\}$

2.4.1 Importance sampling on a path space

In the setting of diffusions, $\mathscr{X} = \mathcal{C}([0,T];\mathbb{R}^d)$, f and g are given by the target and proposal conditioned diffusion measures: $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$ and $\mathrm{d}\mathbb{P}_\mu(\cdot|\mathcal{Z})$ respectively, and the weighting function is proportional to the Radon-Nikodým derivative between the two laws:

$$w(X) \propto \frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X|\mathcal{Z}).$$

Notice that it is no longer necessary to find a global upper bound M^* from C3, which means that A6 and A7 are not the prerequisites for applying the algorithm. Indeed, consider the following example:

Example 2.4.1. Suppose that process X solves the following SDE:

$$\mathrm{d}X_t = \left(b_1 + b_2 X^{[2]}\right)\mathrm{d}t + \begin{pmatrix}\rho & \sqrt{1-\rho^2}\\ \sigma & 0\end{pmatrix}\sqrt{X^{[2]}}\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0,T], \tag{2.16}$$

for some constants $\rho, \sigma > 0$ and $b_1, b_2 \in \mathbb{R}^2$, $b_2^{[1]} = -0.5$. This is the celebrated Heston model (Heston, 1993), in which the first coordinate represents log-price of an asset and the second is the instantaneous volatility of the price process (i.e. of the exponentiated first coordinate). Notice that this is an example of a diffusion for

which Lamperti transformation does not exist—A6 does not hold. Suppose that the conditioning is on the price of an asset to cross an upper threshold $\exp(L^*)$ at some point during time $[0,T]$: $\mathcal{Z} := \{\omega \in \Omega : X_t(\omega) \geq L^* \text{ for some } t \in [0,T]\}$, and that parameters $b_1, b_2, \rho, \sigma, L^*$ are fixed in such a way that $\mathcal{Z}$ is extremely unlikely under $\mathbb{P}_b$.[4] Then, consider a proposal diffusion with the same volatility coefficient and a modified drift:

$$\mu(x) := \mu_1 + \mu_2 x^{[2]},$$

where $\mu_1, \mu_2 \in \mathbb{R}^2$ are some to-be-chosen vectors. The Radon-Nikodým derivative between the target and proposal laws is given by:

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X|\mathcal{Z}) \propto \exp\left\{c_1 + c_2 X_T + c_3 \int_0^T \frac{1}{X_s^{[2]}} I_2 \,\mathrm{d}X_s + c_4 \int_0^T \frac{1}{X_s^{[2]}} \,\mathrm{d}s + c_5 \int_0^T X_s^{[2]} \,\mathrm{d}s\right\}, \tag{2.17}$$

where

$$\Lambda := \frac{1}{(\rho\sigma)^2}\begin{pmatrix} 0 & \rho\sigma \\ \rho\sigma & -1 \end{pmatrix}, \quad c_1 := T(\mu_1^T \Lambda \mu_2 - b_1^T \Lambda b_2), \quad c_2 := (b_2 - \mu_2)^T \Lambda,$$
$$c_3 := (b_1 - \mu_1)^T \Lambda, \quad c_4 := \frac{1}{2}(\mu_1 - b_1)^T \Lambda (b_1 + \mu_1), \quad c_5 := \frac{1}{2}(\mu_2 - b_2)^T \Lambda (b_2 + \mu_2).$$

Proof. Notice

$$\Gamma^{-1}(x) = \frac{1}{x^{[2]}}\Lambda.$$

It follows that:

$$\begin{aligned}\int_0^T [(b-\mu)^T \Gamma^{-1}](X_s)\,\mathrm{d}X_s &= (b_1 - \mu_1)^T \Lambda \int_0^T \frac{1}{X_s^{[2]}} I_2 \,\mathrm{d}X_s + (b_2 - \mu_2)^T \Lambda \int_0^T \mathrm{d}X_s \\ &= (b_1 - \mu_1)^T \Lambda \int_0^T \frac{1}{X_s^{[2]}} I_2 \,\mathrm{d}X_s + (b_2 - \mu_2)^T \Lambda (X_T - X_0)\end{aligned}$$

and

$$\begin{aligned}&\int_0^T [(b-\mu)^T \Gamma^{-1}(b+\mu)](X_s)\,\mathrm{d}s \\ &\qquad = \int_0^T \left\{\frac{1}{X_s^{[2]}}(b_1 - \mu_1)^T \Lambda (b_1 + \mu_1) + (b_1 - \mu_1)^T \Lambda (b_2 + \mu_2)\right. \\ &\qquad\qquad \left. + (b_2 - \mu_2)^T \Lambda (b_1 + \mu_1) + (b_2 - \mu_2)^T \Lambda (b_2 + \mu_2) X_s^{[2]}\right\} \mathrm{d}s\end{aligned}$$

[4] So that simulation from $\mathbb{P}_b(\cdot|\mathcal{Z})$ via simulating unconditioned paths $X \sim \mathbb{P}_b$ and accepting only those for which $\mathcal{Z}$ occurs is an inefficient strategy. See next page for details.

$$= (b_1 - \mu_1)^T \Lambda (b_1 + \mu_1) \int_0^T \frac{1}{X_s^{[2]}} \mathrm{d}s + 2T(b_1^T \Lambda b_2 - \mu_1 \Lambda \mu_2)$$
$$+ (b_2 - \mu_2)^T \Lambda (b_2 + \mu_2) \int_0^T X_s^{[2]} \mathrm{d}s.$$

The statement now follows from eqs. (2.3) and (2.10). □

If vectors μ_1 and μ_2 are chosen carefully enough so that $\mathcal{Z}$ has respectable probability of occurring under the unconditioned proposal measure $\mathbb{P}_\mu$, then samples from $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ can be obtained by drawing $X \sim \mathbb{P}_\mu$ and accepting only those paths for which $\mathcal{Z}$ occurs. In fact, this is a special case of rejection sampling on a path space, in which a proposal law is $\mathbb{P}_\mu$, the target law is $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ and the Radon-Nikodým derivative between the two is proportional to: $\mathrm{d}\mathbb{P}_\mu(X|\mathcal{Z})/\,\mathrm{d}\mathbb{P}_\mu(X) \propto \mathbb{1}_{\mathcal{Z}}$. Draws from $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ may be weighted according to the expression (2.17) in an importance sampling scheme, targeting law $\mathbb{P}_b(\cdot|\mathcal{Z})$. Algorithm 2.7 summarises this procedure.

Algorithm 2.7 Importance sampling for example 2.4.1

1: **for** $i = 1, \ldots, N$ **do**
2: **do**
3: Draw $X_i \sim \mathbb{P}_\mu$
4: **while** $\mathcal{Z}$ does not happen for X_i
5: Set $w_i \leftarrow \exp\left\{c_1 + c_2 X_{i,T} + c_3 \int_0^T \frac{1}{X_{i,s}^{[2]}} I_2 \,\mathrm{d}X_{i,s} + c_4 \int_0^T \frac{1}{X_{i,s}^{[2]}} \mathrm{d}s + c_5 \int_0^T X_{i,s}^{[2]} \mathrm{d}s\right\}$
6: **return** $\{(X_i, w_i); i = 1, \ldots, N\}$

In fig. 2.2 I give the results of a numerical example illustrating implementation of algorithm 2.7.

In spite of what it might seem, importance sampling is not a panacea for the shortcomings of rejection sampling. Despite relaxation of assumptions A6 and A7, care still needs to be taken to avoid nonsensical answers, as although, by construction, importance sampling estimator has a finite mean, there are no such guarantees about its variance. A sufficient (though not necessary) condition implying finiteness of the second moment is condition C2 (Robert and Casella, 2013, §3.3).

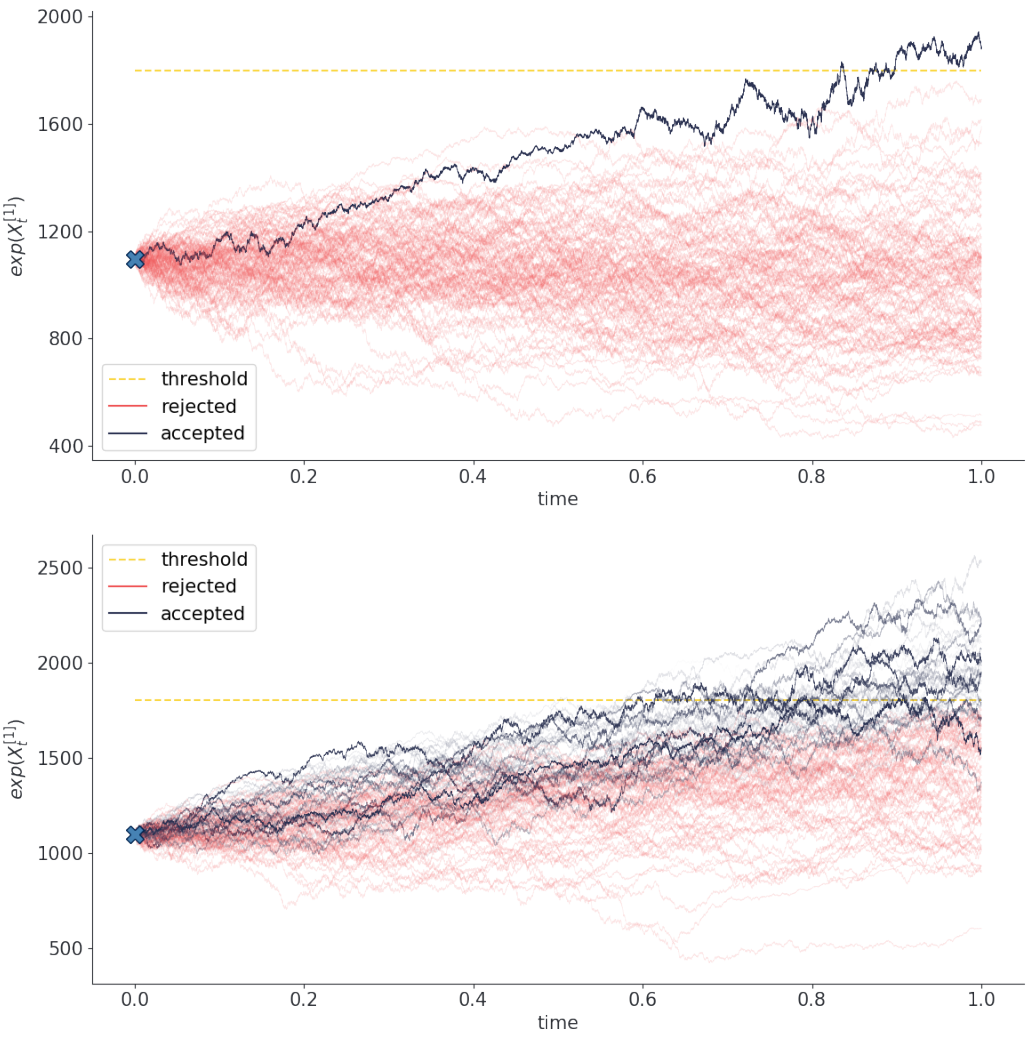


Figure 2.2: Importance sampler on a path space for the Heston model. Underlying path is assumed to follow the dynamics in eq. (2.16), with $b_1 = (0.1, 0.42)^T$, $b_2 = (-0.5, -6)^T$, $\rho = -0.83$ and $\sigma = 0.55$ (where I picked these values to mimic the numerical example considered in Stramer and Bognar (2011)). Diffusion is started from the point $(7, 0.07)^T$ and evolved for a period of time $T = 1$. Threshold L^* is set to 1800. Top plot illustrates the output of a rejection sampler, which draws proposals from an unconditioned target law $\mathbb{P}_b$ and accepts only those paths for which the exponentiated first coordinate exceeds level L^* at some point during $[0, T]$. Because such event $\mathcal{Z}$ is unlikely under $\mathbb{P}_b$ only 1 out of 100 paths has been accepted. Bottom plot illustrates the output of an importance sampling scheme. Paths are drawn from the unconditioned diffusion measure $\mathbb{P}_\mu$, with $\mu_1 = (0.35, 0.42)^T$ and $\mu_2 = (-0.5, -6)$. These paths are first sieved through a rejection sampling step, accepting only those samples for which $\mathcal{Z}$ occurs. Accepted samples are then weighted according to eq. (2.17). The alpha channel of accepted samples is set to be proportional to the value of computed weights. Consequently, the same number of proposals results in a larger number of weighted samples, that can be used to estimate integrals of path functionals. The values of μ_1 and μ_2 were set in an ad-hoc manner, just to illustrate the point, more efficient schemes can be devised with more informed choices of μ_1 and μ_2.

Relying on condition C2 is however not ideal, as it introduces back the need for assumptions A6 and A7. Unfortunately, proving finiteness of variance without relying on condition C2 is in general non-trivial.

2.5 Markov chain Monte Carlo

Importance and rejection sampling alike suffer from a common drawback: there needs to be a way to specify a *good*, global proposal density g. As I discussed in section 2.3.4 this feat is sometimes difficult or even impossible to accomplish. Markov chain Monte Carlo (MCMC) methods address this precise issue, by substituting global update steps with local ones.

Suppose as before that $f : \mathscr{X} \to \mathbb{R}$ is some density function, but that only $f_u : \mathscr{X} \to \mathbb{R}$ can be evaluated, with $f(x) = c_1 f_u(x)$, and $c_1 > 0$ some unknown constant. The goal is to compute integrals of the form:

$$f(\pi) := \int_{\mathscr{X}} \pi(x) f(x) \mathrm{d}x,$$

for f-integrable functions $\pi : \mathscr{X} \to \mathbb{R}$. Markov chain Monte Carlo methods solve this problem by employing f-invariant Markov chains: $\{X^{(n)}; n = 1, \dots\}$. By Ergodic theorem, if such a chain is Harris recurrent, then for any f-integrable function $\pi : \mathscr{X} \to \mathbb{R}$ the strong law of large numbers

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \pi(X^{(n)}) = f(\pi), \quad \text{(a.s.)},$$

holds (Robert and Casella, 2013, Thm 6.63). Consequently, estimators of the form

$$\hat{\pi} := \frac{1}{N} \sum_{i=1}^{N} \pi(X^{(n)}),$$

can be used to approximate $f(\pi)$.

2.5.1 Metropolis-Hastings

The Metropolis-Hastings algorithm due to Metropolis et al. (1953); Hastings (1970) is arguably the most popular out of all MCMC methods (for a justification of this commonly adhered to preference see the results on Peskun ordering (Peskun, 1973)). The chain is initialised at an arbitrary state $X^{(0)}$ and at each new iteration

$n = 1, \ldots, N$ a new value for the chain is proposed by drawing form a user-defined density q:

$$X^\circ \sim q(X^{(n-1)}, \cdot).$$

Proposed value X° is then accepted with the probability $a(X^{(n-1)}, X^\circ)$, where

$$a(x, x^\circ) := 1 \wedge \frac{f_u(x^\circ) q(x^\circ, x)}{f_u(x) q(x, x^\circ)}. \tag{2.18}$$

If accepted, $X^{(n)}$ is set to X°, otherwise, the state is rejected and $X^{(n)}$ is set to $X^{(n-1)}$. Algorithm 2.8 summarises these steps.

Algorithm 2.8 The Metropolis-Hastings algorithm

1: Initialise $X^{(0)}$
2: **for** $n = 0, \ldots, N-1$ **do**
3: Draw $X^\circ \sim q(X^{(n)}, \cdot)$
4: Draw $U \sim \texttt{Unif}([0, 1])$
5: **if** $U < 1 \wedge \frac{f_u(X^\circ) q(X^\circ, X^{(n)})}{f_u(X^{(n)}) q(X^{(n)}, X^\circ)}$ **then**
6: Set $X^{(n+1)} \leftarrow X^\circ$
7: **else**
8: Set $X^{(n+1)} \leftarrow X^{(i)}$
9: **return** $\{X^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant density f

The Metropolis-Hastings algorithm produces an irreducible chain provided q satisfies either of the following two conditions:

Condition C5. *$q(x, x^\circ) > 0$ for all $x, x^\circ \in supp(f)$.*

Condition C6. *(Roberts and Tweedie, 1996) f is bounded on compact sets and there exist $\epsilon > 0$ and $\delta > 0$, for which $q(x, x^\circ) > \epsilon$, whenever $d(x, x^\circ) < \delta$, for some distance metric d.*

In practice, satisfying at least one of C5 or C6 is often simple. If the chain is additionally aperiodic (which, trivially, can be achieved by modifying a transition kernel through introduction of a low-probability move that keeps the chain's state in its current position), by Tierney (1994) the chain is Harris recurrent and since by construction it is also f-invariant, the Ergodic theorem applies.

It is possible to use more than just one transition kernel q. In fact, any finite number of kernels q_k $(k = 1, \ldots, \mathbb{k})$ can be used so long as the resulting chain is

irreducible. At each step n of the MCMC algorithm one of the q_k's is picked and an update is performed by proposing from it. Various strategies for choosing q_k exist. Perhaps the most common ones are updates done according to the lexicographic order—in which q_k's are picked systematically according to the schedule $q_1, \ldots, q_{\mathbb{k}}, q_1, \ldots$—and a random order—in which an index is drawn uniformly at random $k \sim \texttt{Unif}(\{1, \ldots, \mathbb{k}\})$ and then, the corresponding kernel q_k is used to make an update. Different strategies for choosing index k, applied to the same set of kernels $q_1, \ldots, q_{\mathbb{k}}$, may result in a different performance of the MCMC algorithm (Roberts and Sahu, 1997). I discuss this in more detail in the context of blocking in chapter 5.

In this thesis the Metropolis-Hastings algorithm is used in two distinct contexts. By this point, the first use case should be familiar: $\mathscr{X} = \mathcal{C}([0, T]; \mathbb{R}^d)$ and f is given by the target diffusion law $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$. The choice of a proposal kernel q is now less restricted than the choice of a proposal law $\mathrm{d}\mathbb{P}_\mu(\cdot|\mathcal{Z})$ was for rejection or importance sampling. Indeed, as a special case, transition kernel q can be chosen to be $\mathrm{d}\mathbb{P}_\mu(\cdot|\mathcal{Z})$, used in a corresponding rejection or importance sampling scheme—this is summarised in algorithm 2.9. This is a type of the so-called independence sampler, as the proposals X° are drawn from a density which is independent from the previously accepted state $X^{(n-1)}$. However, much more efficient algorithms can be devised if q exploits the possibility of making local updates. I discuss two such strategies in this thesis: the preconditioned Crank-Nicolson scheme, explained in section 2.7 and a blocking scheme, treated in chapter 5. For the latter, in section 5.2, I quantify the computational gains over the corresponding global proposal moves.

Algorithm 2.9 Independence sampler on a path space

1: Draw $X^{(0)} \sim \mathbb{P}_\mu(\cdot|\mathcal{Z})$
2: **for** $n = 0, \ldots, N-1$ **do**
3: Draw $X^\circ \sim \mathbb{P}_\mu(\cdot|\mathcal{Z})$
4: Draw $U \sim \texttt{Unif}([0,1])$
5: **if** $U < 1 \wedge \frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X^\circ|\mathcal{Z}) \Big/ \frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X^{(n)}|\mathcal{Z})$ **then**
6: Set $X^{(n+1)} \leftarrow X^\circ$
7: **else**
8: Set $X^{(n+1)} \leftarrow X^{(n)}$
9: **return** $\{X^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant density $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

The second application has to do with Bayesian inference for diffusion processes. In this context, a Markov chain is constructed to target the posterior distribution of the random and unknown parameter vector $\theta \in \Theta$ parametrising SDE (1.3). The procedure is discussed in detail in chapter 4.

2.5.2 Gibbs sampler

A particular example of the Metropolis-Hastings algorithm is a Gibbs sampler (Geman and Geman, 1987). It is especially useful for sampling from high-dimensional distributions for which coordinates are only weakly correlated. For it, a sequence of transition kernels is defined through unnormalised, conditional target distributions:

$$q_k(x, x^\circ) \propto \mathbb{1}_{\{x^{[-k]}\}}(x^{\circ[-k]}) f_u(x^\circ | x^{[-k]}), \quad k = 1, \ldots, \mathbb{k}.$$

It is not difficult to see that the Metropolis-Hastings acceptance probability (2.18) for each one of those transition kernels equals 1:

$$\begin{aligned} a_k(x, x^\circ) &:= 1 \wedge \frac{f_u(x^\circ) f_u(x | x^{\circ[-k]})}{f_u(x) f_u(x^\circ | x^{[-k]})} = 1 \wedge \frac{f(x^{\circ[k]}, x^{[-k]}) f(x | x^{[-k]})}{f(x) f(x^\circ | x^{[-k]})} \\ &= 1 \wedge \frac{f(x^{\circ[k]}, x^{[-k]}) \frac{f(x)}{f(x^{[-k]})}}{f(x) \frac{f(x^{\circ[k]}, x^{[-k]})}{f(x^{[-k]})}} = 1, \end{aligned}$$

where the second equality follows from $x^{\circ[-k]} = x^{[-k]}$. A Gibbs sampler is defined as the Metropolis-Hastings algorithm which uses the transition kernels above—it is summarised in algorithm 2.10 below.

Algorithm 2.10 Gibbs Sampler with the lexicographic order of coordinate updates

1: Initialise $X^{(0)}$
2: **for** $n = 0, \ldots, N-1$ **do**
3: Set $k \leftarrow (n \mod \mathbb{k}) + 1$
4: Draw $X^\circ \sim f(\cdot | X^{(n)[-k]})$
5: Set $X^{(n+1)} \leftarrow X^\circ$
6: **return** $\{X^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant density f

Instead of the single-site updates, the coordinates can be grouped and updated in blocks instead. More precisely, transition kernels may have the form: $q_{k_1}(x, x^\circ) = f(x^\circ | x^{[-(k_1:k_2)]})$ (where $v^{[-(k_1:k_2)]} := (v_1, \ldots, v_{k_1-1}, v_{k_2+1}, \ldots, v_d)$), so that ranges of coordinates $k_1 : k_2$ are updates simultaneously. This idea can be taken even further.

Example 2.5.1. In the context of sampling on a path space, the object that is to be imputed—path X under some target measure $\mathbb{P}_b(\cdot|\mathcal{Z})$—is conceptually infinite dimensional. Here, block updates take a form of updating segments of path X. In the simplest setting, when $\mathcal{Z} := X_T$ say, I can define two kernels:

$$q_1(X,\cdot) = \mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z} \cup \{X_{T/2}\}), \quad q_2(X,\cdot) = \mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z} \cup \{X_{T/4}, X_{3T/4}\}).$$

Due to the Markov property, the conditional update q_1 updates segments of X independently on each of the intervals $[0, \frac{T}{2}]$ and $[\frac{T}{2}, T]$. Update q_2 does the same, but on the intervals $[0, \frac{T}{4}]$, $[\frac{T}{4}, \frac{3T}{4}]$ and $[\frac{3T}{4}, T]$. In view of the heuristics discussed in section 2.3.4, it is reasonable to expect that such strategy improves the efficiency with which the sampler explores a path space. Detailed analysis of this scheme is given in chapter 5.

In this thesis Gibbs sampling is used extensively for the problem of Bayesian inference for diffusion processes. For numerical experiments, the coordinates of the unknown, random parameter vectors $\theta \in \Theta$ parametrising SDE (1.3) are updated one-by-one or possibly in blocks.

Additionally, a modified version of a Gibbs sampler is often employed—the so-called Metropolis-within-Gibbs algorithm. For it, the transition kernels are allowed to take the form:

$$q_k(x, x^\circ) = \mathbb{1}_{\{x^{[-k]}\}}(x^{\circ[-k]})\tilde{q}_k(x, x^\circ),$$

with $\tilde{q}_k(x, x^\circ)$ possibly different form $f(x^\circ|x^{[-k]})$ (where, as before, single coordinate k can be substituted with a range etc.). Just like with a regular Gibbs sampler, only subsets of all coordinates are updated at a time, unlike it however, the ability to sample from full conditionals is not a prerequisite.

2.6 Non-centred parametrisation

Non-centred parametrisation (Papaspiliopoulos, 2003; Papaspiliopoulos et al., 2003) is an idea of decoupling the source of noise from the specificity of the probability measure under consideration. More precisely, for any collection of probability spaces $(\Omega, \mathscr{F}^{(\theta)}, \mathbb{P}^{(\theta)})_\theta$, indexed by some parameter $\theta \in \Theta$, the aim is to define a

probability space $(\Omega^*, \mathscr{F}^*, \mathbb{Q}^*)$ and a deterministic function $\Psi_\theta : \Omega^* \to \Omega$, so that the pushforward measure $(\mathbb{Q}^*)_{\#}(\Psi_\theta)$ coincides with $\mathbb{P}^{(\theta)}$, i.e. if $W \sim \mathbb{Q}^*$, then $\Psi_\theta(W) \sim \mathbb{P}^{(\theta)}$.

Example 2.6.1. Recall rejection sampler on a path space from example 2.3.1. Suppose that now, the goal is to design a sampler suitable for a sequence of diffusion bridge laws $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$ indexed by $\theta \in \Theta$, with $\mathcal{Z} := X_T$ and with $\mathbb{P}_b^{(\theta)}$ denoting the unconditioned law induced by eq. (1.3). The most straightforward solution is to simply repeat example 2.3.1, each time plugging in a different value of the parameter θ; however, the aim here is to use a non-centred parametrisation—the payoff will be apparent in chapter 4. Assume that A6 and A7 hold. The unconditioned, Lamperti transformed diffusion $Y = \{\eta_\theta(X_t); t \in [0, T]\}$ solves the SDE:

$$\mathrm{d}Y = \alpha_\theta(Y)\,\mathrm{d}t + \mathrm{d}W_t, \quad Y_0 = y_0(\theta), \quad t \in [0, T]. \tag{2.19}$$

Additionally, the conditioning can be transformed to involve only path Y: $\mathcal{Z}_\theta := Y_T = \eta_\theta(X_T)$. Notice that both end-points $y_0(\theta) := \eta_\theta(x_0)$ and $y_T(\theta) := \eta_\theta(x_T)$ depend explicitly on the parameter θ. In example 2.3.1, the randomness enters the sampler in line 2 of algorithm 2.3, when proposal paths are drawn from $\mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z}_\theta)$. This step amounts to drawing paths of Brownian bridges joining $y_0(\theta)$ and $y_T(\theta)$ on the interval $[0, T]$, and thus, because of explicit dependence on θ, it is not of a non-centrally parametrised form. To remedy this, define the following probability space: let $\Omega^* := \mathcal{C}([0, T]; \mathbb{R}^d)$, $\mathbb{Q}^* := \mathbb{W}^*$ be the law induced by Brownian bridges joining 0 and 0 on the interval $[0, T]$ (which I will refer to as 0-0 Brownian bridges) and $\mathscr{F}^*$ be a Borel-σ-algebra on $\mathcal{C}([0, T]; \mathbb{R}^d)$. In simple terms, the noise is sampled by drawing 0–0 Brownian bridges on $[0, T]$. Define also function $\Psi_\theta : \mathscr{X} \to \mathscr{X}$ as:

$$\Psi_\theta(W) := \left\{ W_t + y_0(\theta)\left(1 - \frac{t}{T}\right) + y_T(\theta)\frac{t}{T};\ t \in [0, T] \right\}. \tag{2.20}$$

By eq. (2.13) if $W \sim \mathbb{W}^*$, then $\Psi_\theta(W) \sim \mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z}_\theta)$. Consequently, a version of the non-centrally parametrised rejection sampler on a path space may be defined as given by algorithm 2.11.

This subtle decoupling of a sampling step from the dependence on the parameter θ, in spite of its innocence, has far reaching consequences. Notice that two runs

Algorithm 2.11 Non-centrally parametrised rejection sampling on a path space

1: **while** `True` **do**
2: Draw $W \sim \mathbb{W}^*$ ▷ 0–0 Brownian bridge, use algorithm 2.4
3: Draw $E \sim \mathrm{Exp}(1)$
4: **if** $E \geq \int_0^T \phi_\theta(\Psi_\theta(W)_s)\mathrm{d}s$ **then**
5: Set $X \leftarrow \{\eta_\theta^{-1}(\Psi_\theta(W)_t); t \in [0, T]\}$
6: **return** X ▷ This sample is distributed as $X \sim \mathrm{d}\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$

of algorithm 2.11 parametrised by θ_1 and θ_2 respectively can share the same samples W, entering the algorithm in line 2, without changing the validity of either of the procedures. Such property turns out to be an essential attribute sought after in section 4.1.3.

Non-centred parametrisation is a surprisingly powerful technique that can be used in a number of settings. For instance, it turns out to be an indispensable technique for performing Bayesian inference for diffusion processes via data augmentation (Roberts and Stramer, 2001). A necessity of non-centred parametrisation in this setting is caused by the problem of singularity of measures, discussed further in section 4.1.3. Another application which can be cast in the form of non-centred parametrisation is discussed in section 2.7—a collection Θ consists of a single element and the purpose of employing non-centred parametrisation is to sample the noise on a more convenient probability space. This extra convenience dramatically simplifies the search for efficient local updates of the Metropolis-Hastings algorithm on a path space. A third application is mentioned in section 4.2 and it allows for de-correlating dependence of an auxiliary Poisson point sampler (needed by a rejection sampling scheme) on a model's parameters, ultimately leading to speed-ups of the algorithm.

2.7 Random walk on a path space

The preconditioned Crank-Nicolson scheme introduced in Cotter et al. (2013) (see also Beskos et al. (2008)) can be used in conjunction with the non-centred parametrisation from section 2.6 to perform local moves on a path space under the Metropolis-Hastings algorithm (van der Meulen and Schauer, 2017a). The technique has been

derived from Crank-Nicolson approximations of the linear part of the drift in the context of stochastic partial differential equations (hence its name) but it can be seen simply as a random walk on a path space under the Wiener measure. It is based on the identity:

$$W \overset{d}{=} \sqrt{\lambda} W + \sqrt{1-\lambda} B, \tag{2.21}$$

which holds for any $\lambda \in [0,1]$ and either a pair (W,B) of independent, d-dimensional, standard Brownian motions, or a pair (W,B) of independent, d-dimensional, 0–0 Brownian bridges.

Recall example 2.6.1, treating the non-centred parametrisation of a rejection sampler on a path space. Line 2 of algorithm 2.11 prompts for simulation of Brownian bridges. Instead of drawing them anew at each iteration of the algorithm, eq. (2.21) suggests to use previously sampled Brownian paths. That way, new proposal paths $\Psi_\theta(W)$ should be *close* to the previously accepted ones and thus should also have respectable chance of being accepted. As the proposals are correlated, such algorithm can no longer be phrased in a rejection sampling setting; however, the idea can be rigorously implemented as the Metropolis-Hastings algorithm (which I summarise in algorithm 2.12 below).

Algorithm 2.12 Metropolis-Hastings sampling on a path space with local updates

1: Draw $W^{(0)} \sim \mathbb{W}^*$ ▷ 0–0 Brownian bridge, use algorithm 2.4
2: Set $X^{(0)} \leftarrow \{\eta_\theta^{-1}(\Psi_\theta(W^{(0)})_t); t \in [0,T]\}$
3: **for** $n = 0, \dots, N-1$ **do**
4: Draw $W^\circ \sim \mathbb{W}^*$ ▷ 0–0 Brownian bridge, use algorithm 2.4
5: Set $W^\circ \leftarrow \sqrt{\lambda} W^\circ + \sqrt{1-\lambda} W^{(n)}$
6: Draw $E \sim \mathrm{Exp}(1)$
7: **if** $E \geq \int_0^T \phi_\theta(\Psi_\theta(W^\circ)_s)\,\mathrm{d}s - \int_0^T \phi_\theta(\Psi_\theta(W^{(n)})_s)\,\mathrm{d}s$ **then**
8: Set $(X^{(n+1)}, W^{(n+1)}) \leftarrow (\{\eta_\theta^{-1}(\Psi_\theta(W^\circ)_t); t \in [0,T]\}, W^\circ)$
9: **else**
10: Set $(X^{(n+1)}, W^{(n+1)}) \leftarrow (X^{(n)}, W^{(n)})$
11: **return** $\{X^{(n)}; n = 0, \dots, N\}$ ▷ Markov chain with invariant density $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

Naturally, algorithm 2.12 above is merely an extension of example 2.6.1—it is in particular assumed that A6 and A7 hold and that the conditioning is of the form $\mathcal{Z} := X_T$. Notwithstanding, the preconditioned Crank-Nicolson scheme is a

more general concept that can be used in other settings as well: be it with different forms of conditioning $\mathcal{Z}$ or in conjunction with other proposal laws—guided proposals are one such example (Schauer et al., 2017). In example 4.1.2, following Schauer et al. (2017), I show how to define the non-centrally parametrised guided proposals and then describe how the preconditioned Crank-Nicolson scheme can be employed for them. Additionally, in section 4.1.6 I describe how to use the preconditioned Crank-Nicolson scheme also in the setting of Bayesian inference for diffusions processes.

2.8 Commentary

The number and versatility of the Monte Carlo methods—even upon restricting to only those techniques that are directly relevant to the study of diffusion processes—goes far beyond what can be summarised in a single introductory chapter. The topics that were omitted, but are nonetheless central to a number of successful algorithms for diffusion processes, include sequential Monte Carlo and particle filters (Doucet and Johansen, 2009; Fearnhead et al., 2010), Poisson estimators for estimating density functions (Beskos et al., 2006; Fearnhead et al., 2010), Barker's algorithm (Gonçalves et al., 2017), Monte Carlo Expectation-Maximisation algorithm (Beskos et al., 2006), continuous time sequential importance sampling (Fearnhead et al., 2017), non-reversible Markov chains (Diaconis et al., 2000), pseudo-marginal samplers (Andrieu and Roberts, 2009) and many others.

3

Methods for simulating conditioned diffusions

In chapter 2 I restricted my attention primarly to proposal laws of conditioned Brownian motion $\mathbb{P}_0(\cdot|\mathcal{Z})$. However, the issues discussed in section 2.3.4 regarding discrepancies between proposal and target laws apply not only to rejection sampling on a path space, but to some degree to all Monte Carlo methods. Besides, even though importance sampling and MCMC algorithms on a path space no longer formally require assumptions A6 and A7 to hold, I never introduced any general principles for sampling from proposal laws utilising this extended freedom (I merely considered one example 2.4.1, in which the conditioned-on event $\mathcal{Z}$ is simple enough to rely exclusively on unconditioned samplers). In this chapter I introduce two techniques from statistics literature that use other proposal laws.

The first method is due to Bladt and Sørensen (2014) and Bladt et al. (2016) and in the naming convention of the original paper I refer to it as *simple diffusion bridges*. It is applicable to ergodic, scalar diffusions as well as some multivariate ones. The ingenious design of the proposal law $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ makes $\mathbb{P}_\mu(\cdot|\mathcal{Z})$ approach the target law $\mathbb{P}_b(\cdot|\mathcal{Z})$ as T increases, ultimately rendering this algorithm a fitting complement to the arsenal of methods based on the proposals from $\mathbb{P}_0(\cdot|\mathcal{Z})$.

The second technique—called *guided proposals*—was introduced in Clark (1990) and over the years has been considerably extended by Delyon and Hu (2006); Papaspiliopoulos and Roberts (2012); Schauer et al. (2017); van der Meulen and Schauer (2017a,b, 2018) and Bierkens et al. (2018). It is a versatile and robust instance of the Metropolis-Hastings algorithm, which does not need assumptions A6 and A7 to hold. Even in its simple form its proposal laws $\mathbb{P}_{b^\circ}$ tend to approximate the target $\mathbb{P}_b(\cdot|\mathcal{Z})$ more accurately than conditioned Brownian motion would. However, the algorithm grants an additional, generous degree of flexibility over the choice of the drift coefficient of a proposal law $\mathbb{P}_{b^\circ}$, making it possible to reasonably match the target law even in the multivariate and highly non-linear settings.

Notwithstanding, I start this chapter with another algorithm which aims to accomplish a goal of a different kind, and that is to remove all sources of approximation errors. The result—the exact rejection sampler on a path space due to Beskos

and Roberts (2005); Beskos et al. (2006, 2008)—is an algorithm which outputs independent draws of paths from the target measure $\mathbb{P}_b(\cdot|\mathcal{Z})$, revealed at any user-specified time-grid $(s_1, \ldots, s_M)$ (and some additional, random points) in such a way that no error (beyond that of the floating-point precision of a computer) is present.

At the time of writing this thesis, the publication on simple diffusion bridges contained a subtle mistake. It changes the formulation of the algorithm and hence it is important to be aware of it. In this chapter I present an amended version of Bladt and Sørensen (2014, Theorem 2.1), which corrects this inaccuracy. A corrigendum, written jointly with M. Bladt and M. Sørensen, will be published soon.

The remaining two algorithms are instrumental to further developments of this thesis, as I aim to extend both. In chapter 6 I re-formulate the core computational routines of guided proposals to achieve substantial computational savings, whereas in chapter 7 I extend the exact rejection sampler on a path space, as well as guided proposals to tackle conditioning on first passage times. Additionally, in chapter 5 I develop methods and analyses that address some of the shortcomings of the algorithm of Beskos and Roberts (2005).

I conclude this chapter with a review of other methods from statistics literature for sampling conditioned diffusion bridges. None of the algorithms from this summary are discussed any further in this thesis and thus their descriptions are kept adequately succinct.

3.1 Exact rejection sampling on a path space

Rejection sampling on a path space has been introduced in Beskos and Roberts (2005) and already at its conception an exact version of the algorithm has been formulated. Nonetheless, this thesis separates the exposition of this algorithm into two parts, in order to make a clear distinction between the conceptual formulation of rejection sampling on a path space and the existence of one of its concrete implementations, which does not introduce any approximation errors. The former description has been given in parts of section 2.3, with examples 2.3.1 and 2.3.2 summarising the algorithm. In this section I focus on the second aspect—an implementation of rejection sampling on a path space, that does not introduce any discretisation errors.

The algorithm of Beskos and Roberts (2005) is limited to a class of diffusions which, in addition to A1, A3 A5–A8 and A10 (and if $\mathcal{Z} = \varnothing$ also A11), satisfy the extra assumption A12 below. Other implementations have since been proposed with Beskos et al. (2006) relaxing A12 and Beskos et al. (2008) dispensing with it altogether. In this thesis I discuss the first one and last one of these three settings (i.e. either assuming A12 or dispensing with it altogether). For the former setting, instead of presenting the original construction of Beskos and Roberts (2005), I opt for an—arguably—more elegant version introduced in Beskos et al. (2006). All exact algorithms implementing rejection sampling on a path space described in the literature to this date are limited to $\mathcal{Z} = \varnothing$ or $\mathcal{Z} = X_T$, so throughout this section I assume $\mathcal{Z}$ to be restricted to these two choices. In chapter 7 I introduce a novel extension to first passage time bridges.

Chen and Huang (2013) proposed a variation on the exact implementation of rejection sampling on a path space, applicable to unconditioned diffusions ($\mathcal{Z} = \varnothing$). In spite of it being possible to formally apply the algorithm under weaker conditions than Beskos et al. (2008) (i.e. without assuming A10), a proof of finiteness of expected computational cost of the algorithm is known solely under the assumption A10, just as in Beskos et al. (2008).

In the statistics literature the algorithms of Beskos and Roberts (2005); Beskos et al. (2006, 2008) are known under the names of EA1, EA2 and EA3 respectively (EA stands for Exact Algorithm). I avoid this convention and refer to them explicitly as simple, exact rejection sampling on a path space (for EA1) or exact rejection sampling on a path space via layered construction (for EA3). I use the term exact rejection sampler on a path space to refer generically to all EA-type algorithms.

3.1.1 p-coins for rejection sampler on a path space

Recall examples 2.3.1 and 2.3.2 summarising conceptual rejection sampling on a path space with $\mathcal{Z} = \varnothing$ and $\mathcal{Z} = X_T$. Lines 4 and 5 of algorithm 2.5 are common to both examples and a naïve implementation of this step makes it *the* source of approximation error in both cases. Notice however, that the goal of line 5 of algorithm 2.5 is not to evaluate the integral $\int_0^T \phi(Y_t)\mathrm{d}t$, which is impossible to do

exactly on a computer, but only to draw `Bernoulli(p)`—a Bernoulli random variable with probability of success equal to $\mathtt{p} := \exp\{-\int_0^T \phi(Y_t)\,\mathrm{d}t\}$. This is a general principle often used in this thesis, so it deserves an extra emphasis.

Definition 3.1.1. An algorithm, which for a given $\mathtt{p} \in [0,1]$ outputs iid `Bernoulli(p)` random variables is referred to as a p-coin.

In particular, to draw a p-coin, evaluation of p itself need not be necessary, so long as there exists some procedure which outputs 1 with probability p. For rejection sampling on a path space Beskos et al. (2006) introduced the following idea for a p-coin. Consider a unit intensity Poisson point process $\Phi := \{(\chi_j, \psi_j);\, j = 1, \dots\}$ on $[0,T] \times [0,\infty)$ and denote by $\mathbb{L}$ the law induced by it. Denote also the epigraph of $s \to \phi(Y_s)$ with

$$\mathrm{epi}[\phi(Y)] := \{(s,u) \in [0,T] \times \mathbb{R}_+ : \phi(Y_s) \leq u\},$$

and consider an event

$$\mathscr{E} := \{\omega \in \Omega : \Phi(\omega) \subset \mathrm{epi}[\phi(Y(\omega))]\} = \bigcap_{j \geq 1} \left\{\omega \in \Omega : \phi(Y_{\chi_j(\omega)}(\omega)) < \psi_j(\omega)\right\},$$

that no points of Φ fall below an epigraph of $s \to \phi(Y_s)$. By Devroye (2006, §6), the probability of $\mathscr{E}$ happening under the measure $\mathbb{L}$, conditionally on a realisation of the path Y is given by:

$$\mathbb{L}(\mathscr{E}|Y) = \exp\left\{-\int_0^T \phi(Y_t)\,\mathrm{d}t\right\}. \tag{3.1}$$

Consequently, to sample a p-coin with $\mathtt{p} := \exp\{-\int_0^T \phi(Y_t)\,\mathrm{d}t\}$ for a given path Y, it is enough to draw a Poisson point process Φ of unit intensity on $[0,T] \times [0,\infty)$ and then output 1 if $\mathscr{E}$ occurs.

For a moment, disregard the problem of how to sample a Poisson point process on an infinite slab $[0,T] \times [0,\infty)$—I will discuss it in detail in sections 3.1.2 and 3.1.3—and consider the following modification to algorithm 2.5. Before simulating a trajectory Y in line 3 of algorithm 2.5, sample a Poisson point process Φ with unit intensity on $[0,T] \times [0,\infty)$ and only then, sample Y at times $\{\chi_j; j \geq 1\}$ (coinciding with the times of the Poisson point process). The "if statement" in line 5

of algorithm 2.5 can now be substituted with: "if $\mathscr{E}$ occurs". This modification would have resulted in a valid procedure had the number of Poisson points Φ been finite. Indeed, Y would have been needed to be revealed only at a finite collection of random times: $\{\chi_j;\, j=1,\dots,\varkappa\}$ (where I use the convention $\{\cdot;\, j=1,\dots,0\}:=\varnothing$) and the "if statement" would have been reduced to checking if $\phi(Y_{\chi_j})<\psi_j$ were true for all $j=1,\dots,\varkappa$. The accepted path Y (revealed at $\{\chi_j;\, j=1,\dots,\varkappa\}$) would have been distributed exactly according to $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$. The following result makes the last statement rigorous.

Theorem 3.1.1. (Beskos et al., 2008, Theorem 1) If $(Y,\Phi)\sim[\mathbb{P}_0\otimes\mathbb{L}](\cdot|\mathcal{Z},\mathscr{E})$, the marginal distribution of Y is given by $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$. An analogous statement holds true for the unconditioned diffusions: if $(Y,\Phi)\sim[\mathbb{Z}\otimes\mathbb{L}](\cdot|\mathscr{E})$ the marginal distribution of Y is given by $\mathbb{P}_\alpha$.

Proof. If $(Y,\Phi)\sim[\mathbb{P}_0\otimes\mathbb{L}](\cdot|\mathcal{Z},\mathscr{E})$, the marginal distribution of Y can be denoted as:

$$\mathrm{d}\int[\mathbb{P}_0\otimes\mathbb{L}](Y,\Phi|\mathcal{Z},\mathscr{E})\,\mathrm{d}\Phi. \tag{3.2}$$

Since $\mathbb{P}_0(\cdot|\mathcal{Z})$ is a dominating measure for both (3.2) and $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$ it is enough to show that

$$\frac{\mathrm{d}\int[\mathbb{P}_0\otimes\mathbb{L}](Y,\Phi|\mathcal{Z},\mathscr{E})\,\mathrm{d}\Phi}{\mathrm{d}\mathbb{P}_0(Y|\mathcal{Z})}\propto\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y|\mathcal{Z}).$$

This statement follows easily from Bayes' theorem and eqs. (2.11) and (3.1) (eq. (2.14) in place of eq. (2.11) for unconditioned diffusions):

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y|\mathcal{Z})&\propto\exp\left\{-\int_0^T\phi(Y_t)\,\mathrm{d}t\right\}=\mathbb{L}(\mathscr{E}|Y)\\&=\frac{\mathrm{d}\int[\mathbb{P}_0\otimes\mathbb{L}](Y,\Phi|\mathcal{Z})\mathbb{I}_{\mathscr{E}}(Y,\Phi)\,\mathrm{d}\Phi}{\mathrm{d}\mathbb{P}_0(Y|\mathcal{Z})}\\&=\frac{\mathrm{d}\int[\mathbb{P}_0\otimes\mathbb{L}](Y,\Phi|\mathcal{Z},\mathscr{E})\,\mathrm{d}\Phi}{\mathrm{d}\mathbb{P}_0(Y|\mathcal{Z})}[\mathbb{P}_0\otimes\mathbb{L}](\mathscr{E}|\mathcal{Z})\\&\propto\frac{\mathrm{d}\int[\mathbb{P}_0\otimes\mathbb{L}](Y,\Phi|\mathcal{Z},\mathscr{E})\,\mathrm{d}\Phi}{\mathrm{d}\mathbb{P}_0(Y|\mathcal{Z})}.\end{aligned}$$

The proof for unconditioned diffusions follows similarly. □

Additionally, after accepting path Y (which post-acceptance would have been revealed only at a finite and random collection of time points), it would have been possible to retrospectively reveal it at any other collection of time points by simply sampling from the proposal measure $\mathbb{P}_0(\cdot|\mathcal{Z})$, conditionally on all the points that would have been simulated already $\{Y_{\chi_j}; j = 1,\dots,\varkappa\}$—i.e. by sampling from the laws of d-dimensional Brownian bridges.

Nonetheless, this algorithm remains hypothetical until I show how to simulate a Poisson point process on an infinite slab $[0,T]\times[0,\infty)$ in such a way that only a finite number of points is relevant. This is the subject of the next two sections delineating methods that achieve this feat.

3.1.2 Simple, exact rejection sampling on a path space

The simplest way of making the hypothetical algorithm above practically possible is to assume A12 below holds.

Assumption A12. *There exists a constant $l^* > \infty$, s.t. $l^* \geq \sup\{\phi(x); x \in \mathbb{R}^d\}$.*

The algorithm of Beskos and Roberts (2005) rests on this assumption (note that Beskos and Roberts (2005) introduced a different p-coin; the p-coin described above is due to Beskos et al. (2006)). Notice that under this condition any Poisson point simulated on $[0,T]\times(l^*,\infty)$ must automatically fall on the epigraph of $s \to \phi(Y_s)$. Consequently, there is no reason to simulate those points and instead simulating Poisson point process Φ on $[0,T]\times[0,l^*]$ is sufficient (and $|\Phi| := \varkappa < \infty$ almost surely). This results in the following algorithm 3.1.

Algorithm 3.1 Simple, exact rejection sampling on a path space

1: **while** `True` **do**
2: Draw $\Phi \sim \mathbb{L}$ on $[0,T]\times[0,l^*]$ ▷ See Devroye (2006, §VI.1)
3: Draw $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z})$ at times $\{\chi_j; j = 1,\dots,\varkappa\}$ ▷ Use algorithm 2.4
4: **if** $\phi(Y_{\chi_j}) < \psi_j$ for all $j = 1,\dots,\varkappa$ **then**
5: Draw $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z},\{Y_{\chi_j}; j = 1,\dots,\varkappa\})$ ▷ Reveal path at additional times
6: Set $X \leftarrow \{\eta^{-1}(Y_t); t \in [0,T]\}$
7: **return** X ▷ This sample is distributed as $X \sim \mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

As a remark, a non-centrally parametrised version of this algorithm would substitute line 3 with simulation of $W \sim \mathbb{W}^*$ and subsequent $Y \leftarrow \Psi_\theta(W)$ (with Ψ_θ defined in eq. (2.20)).

3.1.3 Layered construction of Brownian bridges

Although assumption A12 can always be artificially enforced by modifying the coefficients of an SDE (truncating them at some large enough hypercube) this can result in a very large upper bound l^* and thus expensive simulations; additionally such amendment introduces bias. Consequently, removing A12 is desirable. To this end, Beskos et al. (2008) proposed the following modification based on an ingenious construction of Brownian bridges.

Suppose it were possible to simulate bounds on the minimum and maximum of each coordinate of d-dimensional Brownian bridges and then, simulate d-dimensional Brownian bridges conditionally on the minimum and maximum bounds on each of their coordinates (and conditionally on any other random variables simulated on the way). By A8, function ϕ is bounded on compact intervals (for any compact set $C \in \mathbb{R}^d$: $\sup_{y \in C} \phi(y) \leq l^*(C) < \infty$) and thus, the following modification to algorithm 3.1 can be employed. First, simulate a hypercube which contains proposal path Y. Use Υ to denote a collection of all random variables that were needed to be simulated for it. Then, local upper bound $l^*(\Upsilon)$ can be computed and the algorithm can proceed as before. Optionally, local lower bound $l_*(\Upsilon)$ can also be computed allowing for an employment of a preliminary rejection step, which slightly accelerates the procedure, but for clarity of exposition I omit this extension (see Pollock (2013)). This is summarised in algorithm 3.2. It remains to show how to execute lines 2, 5 and 7.

Beskos et al. (2008) propose to use the layered construction of Brownian bridges. It starts with fixing an increasing sequence: $0 = a_0 < a_i < \ldots$ with $a_i \uparrow \infty$, and defining the random variable $\mathscr{I}$ (taking values in $\mathbb{N}_+^d$), via:

$$\{\mathscr{I}^{[j]} \leq i\} := \{Y_t^{[j]} \in [y_0^{[j]} \wedge y_T^{[j]} - a_i, y_0^{[j]} \vee y_T^{[j]} + a_i];\, t \in [0,T]\}, \quad i = i \geq 1, \quad j = 1,\ldots,d. \tag{3.3}$$

As the coordinates of the process Y are independent, so are the coordinates of the random variable $\mathscr{I}$. Equation (3.3) says that $\mathscr{I}^{[j]}$ takes a value smaller or equal to

Algorithm 3.2 Exact rejection sampling on a path-space via local bounds

while `True` **do**
  Draw Υ ▷ see below
  Compute $l^*(\Upsilon)$
  Draw $\Phi \sim \mathbb{L}$ on $[0,T]\times[0,l^*(\Upsilon)]$ ▷ See Devroye (2006, §VI.1)
  Draw $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z},\Upsilon)$ at times $\{\chi_j;\, j=1,\dots,\varkappa\}$ ▷ see below
  if $\phi(Y_{\chi_j}) < \psi_j$ for all $j=1,\dots,\varkappa$ **then**
    Draw $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z},\Upsilon,\{Y_{\chi_j};\, j=1,\dots,\varkappa\})$ ▷ see below
    Set $X \leftarrow \{\eta^{-1}(Y_t); t\in[0,T]\}$
    return X ▷ This sample is distributed as $X \sim \mathrm{d}\mathbb{P}_b$

i, if the path $Y^{[j]}$ stays on the interior of the interval $[y_0^{[j]}\wedge y_T^{[j]}-a_i, y_0^{[j]}\vee y_T^{[j]}+a_i]$. Consequently, $\mathscr{I}$ can be used to find a hypercube bounding path Y. Elementary set manipulations show that

$$\{\mathscr{I}^{[j]}=i\} = U_i^{[j]}\cup L_i^{[j]}, \quad i=1\geq 1, \quad j=1,\dots,d,$$

where

$$U_i^{[j]} = \left\{\sup_{0\leq s\leq T} Y_s^{[j]} \in [y_0^{[j]}\vee y_T^{[j]}+a_{i-1}, y_0^{[j]}\vee y_T^{[j]}+a_i)\right\}\cap\left\{\inf_{0\leq s\leq T} Y_s^{[j]} > y_0^{[j]}\wedge y_T^{[j]}-a_i\right\},$$

$$L_i^{[j]} = \left\{\inf_{0\leq s\leq T} Y_s^{[j]} \in (y_0^{[j]}\wedge y_T^{[j]}-a_i, y_0^{[j]}\wedge y_T^{[j]}-a_{i-1}]\right\}\cap\left\{\sup_{0\leq s\leq T} Y_s^{[j]} < y_0^{[j]}\vee y_T^{[j]}+a_i\right\},$$

$$i\geq 1, \quad j=1,\dots,d.$$

Notice that $U_i^{[j]}$ says simply that the maximum value reached by the j^{th} coordinate of the process Y belongs to the interval $[y_0^{[j]}\vee y_T^{[j]}+a_{i-1}, y_0^{[j]}\vee y_T^{[j]}+a_i)$ and that the process is also bounded from below by $y_0^{[j]}\wedge y_T^{[j]}-a_i$ on $[0,T]$. Similar reasoning applies to $L_i^{[j]}$. This is depicted graphically in fig. 3.1. For obvious reasons $\mathscr{I}$ is referred to as a layer or a layer information.

Simulation of $\mathscr{I}$ and the subsequent steps of sampling from $\mathbb{P}_0(\cdot|\mathcal{Z},\mathscr{I})$ and $\mathbb{P}_0(\cdot|\mathcal{Z},\mathscr{I},\{\chi_j;\, j=1,\dots,\varkappa\})$ are quite involved (they are based on identities from Wang and Pötzelberger (1997) and a series method for simulating p-coins from Devroye (2006, §IV.5)), their details are not of the utmost importance to the results of this thesis and are thus omitted (see Beskos et al. (2008) for details). With these samplers in place, Beskos et al. (2008) set $\Upsilon := \mathscr{I}$ and use algorithm 3.2 to perform

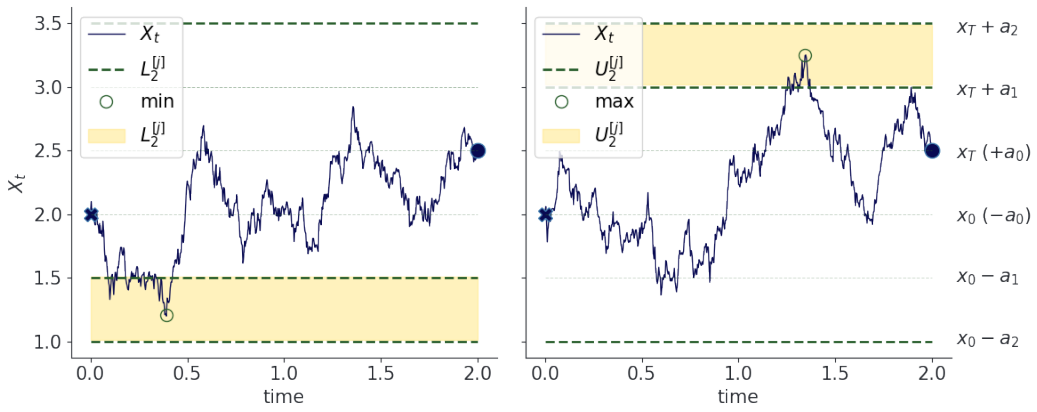


Figure 3.1: Illustration of the 2^{nd} layer on the j^{th} coordinate of a d-dimensional Brownian bridge. $L_2^{[j]}$'th layer is given on the left and $U_2^{[j]}$'th layer on the right. The path of a Brownian bridge, drawn here for illustrative purpose, remains latent in practice.

exact rejection sampling on a path space. As a remark, a non-centred parametrisation of algorithm 3.2 follows easily by drawing layer information $\mathscr{I}$ for the process $W \sim \mathbb{W}^*$ (a d-dimensional, 0–0 Brownian bridge) instead of $Y \sim \mathbb{P}_0$, setting $Y \leftarrow \Psi_\theta(W)$ and computing a hypercube bounding Y from a hypercube bounding W.

3.1.4 Computational cost

As discussed in section 2.3.4, one of the problems afflicting rejection sampling on a path space is the unfavourable scaling of its computational cost with T (the duration of the bridge). This is the central motivation for some of the developments of this thesis, discussed in chapter 5. In this section I will present a formal argument justifying a commonly repeated claim that the computational cost of the exact rejection sampler on a path space scales exponentially with T.

First, recall that a proposal path $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z})$ is accepted with the probability given by eq. (2.11) (so for a given Y the number of trials until the first acceptance is distributed as a geometric random variable with mean given by the reciprocal of eq. (2.11)). The computational cost of each proposal draw is proportional to the number of simulated Poisson points, and for any given Y the expected number of such points is equal to $l^*(\Upsilon)T$. Denote by c_1 the average computational complexity associated with a single Poisson point—I assume all other costs to be negligible.

The expected computational complexity of each call to rejection sampler on a path space is therefore given by

$$\mathscr{R} := \mathbb{E}_{\Upsilon}\left[\mathbb{E}_{Y}\left[c_1 l^*(\Upsilon) T \exp\left\{\int_0^T (\varphi(Y_s) - l_*)\,\mathrm{d}s\right\}\middle|\,\Upsilon\right]\right],$$

where the subscript in the expectation indicates a random variable being integrated out. In simple settings this expression can be used to find an asymptotic lower bound for $\mathscr{R}$ as $T \to \infty$.

Suppose that l^* is independent of Υ (i.e. that assumption A12 holds). If l_* is strictly smaller than the infimum of φ, then there exists $c_2 > 0$, s.t. $\varphi(y) - l_* > c_2$ for all $y \in \mathscr{X}$, and thus

$$\mathscr{R} = c_1 l^* \mathbb{E}\left[\exp\left\{\int_0^T (\varphi(Y_s) - l_*)\,\mathrm{d}s\right\}\right] \geq c_1 l^* T \exp\{c_2 T\}.$$

Typically, even if l_* is the minimum of φ, i.e. if $l_* = \inf_{y\in\mathscr{X}} \varphi(y) = \varphi(\check{m})$, then it is possible to find an ϵ-ball $B_\epsilon(\check{m})$ around $\check{m}$ such that $\varphi(y) - l_* > c_2$ holds for all $y \in [B_\epsilon(\check{m})]^C$. If the end-points of Y belong to $[B_\epsilon(\check{m})]^C$ it is easy to convince oneself that for any $s \in [0, T]$: $\mathbb{P}_0(Y_s \in [B_\epsilon(\check{m})]^C | \mathcal{Z}) > c_3$ for some constant $c_3 := c_3(Y_0, Y_T) > 0$ and T large enough. It then follows by Jensen's inequality and Fubini's theorem:

$$\begin{aligned}
\mathscr{R} &\geq c_1 l^* T \mathbb{E}\left[\exp\left\{\int_0^T (\varphi(Y_s) - l_*)\mathbb{1}_{[B_\epsilon(\check{m})]^C}(Y_s)\,\mathrm{d}s\right\}\right] \\
&\geq c_1 l^* T \exp\left\{\int_0^T c_2 \mathbb{P}_0\left[Y_s \in [B_\epsilon(\check{m})]^C\right]\mathrm{d}s\right\} \\
&\geq c_1 l^* T \exp\{c_2 c_3 T\}.
\end{aligned}$$

Additionally, under A12 the upper bound on $\mathscr{R}$ follows immediately:

$$\mathscr{R} \leq c_1 l^* T \exp\{l^* T\}.$$

The two bounds above, sandwiching $\mathscr{R}$, justify the claim that a computational cost of exact rejection sampling on a path space scales exponentially with T.

3.2 Simple Diffusion Bridges

Simple diffusion bridges, introduced in Bladt and Sørensen (2014), work with unparalleled efficiency at simulating scalar, ergodic diffusion bridges over long time-intervals. The method has also been extended to multivariate settings in Bladt et al.

(2016); however, it is beyond the scope of this thesis, so I restrict the summary below to the results from Bladt and Sørensen (2014) only. In particular $\mathcal{Z} := X_T$ is assumed throughout.

The bridges are constructed in an unorthodox manner. The first curiosity comes from the fact that the algorithm only ever forward simulates unconditioned diffusions. Such paths of unconditioned diffusions are then carefully *spliced* together at a well specified crossing time to form a single proposal path. The proposals are subsequently embedded in the Metropolis-Hastings algorithm, targeting $\mathbb{P}_b(\cdot|\mathcal{Z})$, and the "accept/reject" step is completed through the means of simulating additional, auxiliary unconditioned diffusions. This method has two convincing advantages: it uses simpler samplers of unconditioned diffusions, for which many highly efficient and mature algorithms exist (Kloeden and Platen, 2013) and additionally, the longer the time interval $[0, T]$ is, the more efficient the method becomes.

Throughout, assume that A13 below holds

Assumption A13. *The density of the speed measure*

$$m(x) := \frac{1}{\sigma^2(x)} \exp\left\{2\int_0^x \frac{b(u)}{\sigma^2(u)}\,\mathrm{d}u\right\}, \quad x \in \mathbb{R},$$

of diffusion X solving (1.1) *is finite: $\int_{\mathbb{R}} m(x)\,\mathrm{d}x < \infty$. I.e. diffusion X is ergodic.*

Denote with ν the invariant measure of X and consider three independent diffusions $X^{(i)}$, $i = 1, 2, 3$, solutions to (1.1) on $[0, T]$ (each driven by independent Brownian motion), that are conditioned to start from $X_0^{(1)} = x_0$, $X_0^{(2)} = x_T$, and $X_0^{(3)} \sim \nu(\cdot)$, respectively. I refer to the last of these three as the *auxiliary diffusion*. Define further $\overleftarrow{X}^{(2)} := \overleftarrow{X}^{T,(2)} := \{X^{(2)}_{T-t},\, t \in [0, T]\}$, $\tau^{(Z)} := \inf\{0 \le t \le T : X_t^{(1)} = \overleftarrow{X}_t^{(2)}\}$ and

$$Z_t := \begin{cases} X_t^{(1)} & \text{if } 0 \le t \le \tau^{(Z)}, \\ \overleftarrow{X}_t^{(2)} & \text{if } \tau^{(Z)} < t \le T. \end{cases} \tag{3.4}$$

Let $\mathscr{E}_x$ denote the set of all functions $y \in \mathcal{C}([0, T]; \mathbb{R})$ that intersect $x \in \mathcal{C}([0, T]; \mathbb{R})$:

$$\mathscr{E}_x := \{y \in \mathcal{C}([0, T]; \mathbb{R}) | \mathrm{gr}(y) \cap \mathrm{gr}(x) \neq \varnothing\}, \tag{3.5}$$

where $\mathrm{gr}(\cdot)$ denotes a graph of a function—i.e. $\mathrm{gr}(f) = \{(t, f(t)) | t \in [0, T]\}$ for a given function $f : t \mapsto f(t)$. Then the following can be shown:

Theorem 3.2.1 (Corrected Theorem 2.1 from Bladt and Sørensen (2014))**.** The following equivalence in distribution holds:

$$Z \,\big|\, \{\tau^{(Z)} \leq T\} \overset{d}{=} X \,\big|\, \{X_0 = x_0, X_T = x_T, X^{(3)} \in \mathscr{E}_X\}. \tag{3.6}$$

In words, path Z is distributed as a target diffusion bridge $X \sim \mathbb{P}_b(\cdot|\mathcal{Z})$ additionally conditioned on being hit by an independent diffusion with the unconditioned target dynamics $\mathbb{P}_b$, started from an invariant distribution ν.

Proof. Define the following stopping time:

$$\rho := \rho(X_t^{(1)}, X_t^{(3)}) := \inf\{t \geq 0 : X_t^{(1)} = X_t^{(3)}\},$$

and the process Y by:

$$\begin{cases} Y_t := \begin{cases} X_t^{(1)} & \text{if } 0 \leq t \leq \rho, \\ X_t^{(3)} & \text{if } \rho < t \leq T, \end{cases} & \text{if } \rho \leq T, \\ Y_t := X_t^{(1)}, & \text{otherwise.} \end{cases}$$

Notice that by the definition of Y: $Y_T = X_T^{(3)}$ on $\{\rho \leq T\}$, and thus

$$Y \,\big|\, \{Y_T = x_T, \rho \leq T\} \overset{d}{=} Y \,\big|\, \{X_T^{(3)} = x_T, \rho \leq T\}. \tag{3.7}$$

Additionally, Bladt and Sørensen (2014, Lemma 2.2) prove that

$$X^{(3)} \,\big|\, \{X_T^{(3)} = x_T\} \overset{d}{=} \overleftarrow{X}^{(2)},$$

from which it follows that

$$Y \,\big|\, \{X_t^{(3)} = x_T, \rho \leq T\} \overset{d}{=} Z \,\big|\, \{\tau^{(Z)} \leq T\}. \tag{3.8}$$

Combining eqs. (3.7) and (3.8) yields

$$Y \,\big|\, \{Y_T = x_T, \rho \leq T\} \overset{d}{=} Z \,\big|\, \{\tau^{(Z)} \leq T\}. \tag{3.9}$$

Now, by the strong Markov property Y has the same distribution as $X^{(1)}$, and the same holds true on $\{\rho \leq T\}$, i.e.:

$$Y \,\big|\, \{\rho \leq T\} \overset{d}{=} X^{(1)} \,\big|\, \{\rho \leq T\}. \tag{3.10}$$

In eq. (3.10) above I can now condition on the end-point of the process whose distribution is being considered, i.e. Y_T on the left hand side and $X_T^{(1)}$ on the right:

$$Y \,\big|\, \{Y_T = x_T, \rho \leq T\} \overset{d}{=} X^{(1)} \,\big|\, \{X_T^{(1)} = x_T, \rho \leq T\}. \tag{3.11}$$

Combining eq. (3.11) with eq. (3.9) yields:

$$Z \,\big|\, \{\tau^{(Z)} \leq T\} \overset{d}{=} X^{(1)} \,\big|\, \{X_T^{(1)} = x_T, \rho \leq T\}, \tag{3.12}$$

which is precisely the statement of the theorem. □

Remark 3.2.1. The original statement of Bladt and Sørensen (2014, Theorem 2.1) read

$$Z \,\big|\, \{\tau^{(Z)} \leq T\} \overset{d}{=} X \,\big|\, \{X_0 = x_0, X_T = x_T, X^{(4)} \in \mathscr{E}_X\},$$

where $X^{(4)}$ was a different auxiliary diffusion, defined to follow the dynamics (1.1) on $[0, T]$ and taken to start from a random point with distribution $X_0 \sim p_T(x_T, \cdot)$, where p_t denotes the transition density of X. This conclusion was a result of the erroneous statement of eq. (3.11), which in the paper was assumed to say:

$$Y \,\big|\, \{Y_T = x_T, \rho \leq T\} \overset{d}{=} X^{(1)} \,\big|\, \{X_T^{(1)} = x_T, \boldsymbol{X_T^{(3)} = x_T}, \rho \leq T\}.$$

The same correction extends to Bladt et al. (2016, Theorem 2.2). An important consequence of this amendment is that in order to use the algorithms of Bladt and Sørensen (2014) and Bladt et al. (2016) it is not only enough for the invariant density ν to exist, but it must also be possible to sample from it. A corrigendum written jointly with M. Bladt and M. Sørensen will be published shortly.

Using the last, summarising sentence of theorem 3.2.1 it follows that I can denote the law of the process $Z|\{\tau^{(Z)} \leq T\}$ with $\mathbb{P}_b(\cdot|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})$. This law can be chosen to be a proposal law for an MCMC sampler targeting the distribution $\mathbb{P}_b(\cdot|\mathcal{Z})$.

To sample from $\mathbb{P}_b(\cdot|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})$ a simple rejection sampling algorithm can be employed. The pairs $(X^{(1)}, X^{(2)})$ are generated until the first occurrence of $\{\tau^{(Z)} \leq T\}$, upon which Z is returned. Algorithm 3.3 summarises this sampling procedure.

Algorithm 3.3 Sampler of proposal bridges for `SDB`

Set $(X_0^{(1)}, X_0^{(2)}) \leftarrow (x_0, x_T)$
while True **do**
  Sample path $X^{(1)} \sim \mathbb{P}_b$ on $[0,T]$
  Sample path $X^{(2)} \sim \mathbb{P}_b$ on $[0,T]$
  Set $\overleftarrow{X}^{(2)} \leftarrow \{X^{(2)}_{T-t};\, t \in [0,T]\}$
  Set $\tau^{(Z)} \leftarrow \inf\{t \geq 0 : X_t^{(1)} = \overleftarrow{X}_t^{(2)}\}$
  if $\tau^{(Z)} < \infty$ **then**
    Set $Z \leftarrow \{X_t^{(1)} \mathbb{1}_{\{t \leq \tau^{(Z)}\}} + \overleftarrow{X}_t^{(2)} \mathbb{1}_{\{t > \tau^{(Z)}\}};\, t \in [0,T]\}$
    return Z ▷ Path distributed as $\mathbb{P}_b(\cdot|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})$

To sample $(X^{(1)}, X^{(2)})$ Bladt and Sørensen (2014) use discretisation schemes based on stochastic Taylor expansions (Kloeden and Platen, 2013). Correspondingly, they determine if $\tau^{(Z)}$ occurred by linearly interpolating the diffusion path between the values taken on a sampled time-grid.

The Radon-Nikodým derivative between the target and the proposal diffusion laws follows easily from Bayes' formula:

$$\frac{\mathrm{d}\mathbb{P}_b(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_b(X|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})} = \frac{\mathbb{P}_b((X, X^{(3)}) \in \mathscr{E}|\mathcal{Z})}{\mathbb{P}_b(X^{(3)} \in \mathscr{E}_X|X)},$$

where

$$\mathscr{E} := \{(x,y) \in \mathcal{C}([0,T];\mathbb{R}) \times \mathcal{C}([0,T];\mathbb{R}) | y \in \mathscr{E}_x\}.$$

Proof. Notice:

$$\mathrm{d}\mathbb{P}_b(X|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\}) = \mathbb{P}_b(X^{(3)} \in \mathscr{E}_X|\mathcal{Z}, X) \frac{\mathrm{d}\mathbb{P}_b(X|\mathcal{Z})}{\mathbb{P}_b(X^{(3)} \in \mathscr{E}_X|\mathcal{Z})}.$$

The statement follows from basic algebra and after realising that $\mathcal{Z}$ can be dropped from the conditioning in $\mathbb{P}_b(X^{(3)} \in \mathscr{E}_X|\mathcal{Z}, X)$. □

Define:

$$\pi(Z) := \mathbb{P}_b(X^{(3)} \in \mathscr{E}_Z|Z),$$

and notice that the Metropolis-Hastings algorithm which uses $Z \sim \mathbb{P}_b(\cdot|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})$ to target $\mathbb{P}_b(\cdot|\mathcal{Z})$ has acceptance probability of the form:

$$a(Z^{(n)}, Z) = 1 \wedge \frac{\pi(Z^{(n)})}{\pi(Z)}.$$

This expression cannot be evaluated in a closed form; however, it is possible to obtain a positive, unbiased estimator of $1/\pi(Z)$. Such estimator can be simply plugged in for a value of $1/\pi(Z)$ and the invariant distribution of the resulting Markov Chain is unaltered (Andrieu and Roberts, 2009). This is an example of a *pseudo-marginal MCMC*.

For the proposal sample $Z \in \mathcal{C}([0,T];\mathbb{R})$ one such estimator can be defined as

$$\mathscr{T} := \min\{i \in \mathbb{N}_+ : X^{(3,i)} \in \mathscr{E}_Z\},$$

with $X^{(3,i)}$, $i \in \mathbb{N}_+$, iid copies of the auxiliary diffusion $X^{(3)}$ ($\mathscr{T}$ is just a Geometric random variable with probability of success $\pi(Z)$, and thus $\mathbb{E}[\mathscr{T}] = 1/\pi(Z)$). A lower variance estimator (coming at a higher computational cost) is given by an average of independent realisations of $\mathscr{T}$:

$$\overline{\mathscr{T}} := \frac{1}{N}\sum_{j=1}^{N} \mathscr{T}_j$$

The resulting MCMC sampler is given in algorithm 3.4.

Algorithm 3.4 Simple diffusion bridges (`SDB`; pseudo-marginal independence sampler)

Draw $Z^{(0)} \sim \mathbb{P}_b(\cdot|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})$ ▷ Use algorithm 3.3
Set $\mathscr{T}^{(0)} \leftarrow 0$
for $n = 1, \dots, N$ **do**
  Draw $Z \sim \mathbb{P}_b(\cdot|\mathcal{Z}, \{X^{(3)} \in \mathscr{E}_X\})$ ▷ Use algorithm 3.3
  Set $\mathscr{T} \leftarrow 0$
  repeat
    $\mathscr{T} \leftarrow \mathscr{T} + 1$
    Sample $X_0^{(3)} \sim \nu$ ▷ Sampling from the invariant density of $\mathbb{P}_b$
    Sample $X^{(3)} \sim \mathbb{P}_b$ on $[0,T]$
    Set $\tau^{(\mathrm{aux})} \leftarrow \inf\{t \geq 0 : X_t^{(3)} = Z_t\}$ ▷ With the convention $\inf\varnothing := \infty$
  until $\tau^{(\mathrm{aux})} < \infty$
  Draw $U \sim$ `Unif([0,1])`
  if $U \leq \mathscr{T}/\mathscr{T}^{(n)}$ **then**
    Set $(Z^{(n)}, \mathscr{T}^{(n)}) \leftarrow (Z, \mathscr{T})$
  else
    Set $(Z^{(n)}, \mathscr{T}^{(n)}) \leftarrow (Z^{(n-1)}, \mathscr{T}^{(n-1)})$
return $\{Z^{(n)}; n = 0, \dots, N)$ ▷ Chain with the invariant density $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

Bladt and Sørensen (2014) once again rely on discretisations to simulate $X^{(3,i)}$ and they verify occurrence of $\tau^{(\mathrm{aux})}$ accordingly.

3.3 Guided Proposals

Guided proposals is a Markov chain Monte Carlo algorithm on a path space, designed around the premise of making only the most practical assumptions about the underlying diffusion process. In particular, the restrictive assumptions A6 and A7 do not need to hold, nor is it necessary for the underlying diffusion to be ergodic. The first seed of this method can be found in Clark (1990), where the simplest, one dimensional setting was considered with the diffusion X solving eq. (1.1) having the volatility coefficient set to a constant. Delyon and Hu (2006) substantially extended this preliminary work to multivariate settings, diffusions with general volatility coefficients and even some limited, hypoelliptic diffusions. Furthermore, Papaspiliopoulos and Roberts (2012) derived the expression for the normalisation constants that are required for applying the algorithm of Delyon and Hu (2006) to inference; they were also the first ones to have coined the term 'guided proposals'. Additionally, the work of Durham and Gallant (2002), set in a discrete-time setting, bears connections to the work of Delyon and Hu (2006); however the connection itself has not been made clear until later (see section 3.4 for details).

More recently, Schauer et al. (2017) introduced a further generalisation, which dispenses with invertibility of the volatility coefficient and grants a generous degree of flexibility over the choice of a proposal process, often bringing simulation of radically non-linear diffusions into the realm of feasible problems. In van der Meulen and Schauer (2017b) and van der Meulen and Schauer (2018) guided proposals have been further extended to the conditioning on the partial observations of the underlying process with additive noise:

$$\mathcal{Z} := \{L_i X_{t_i}(\omega) + \xi_i(\omega),\, i = 1,\ldots,K\}, \tag{3.13}$$

with $L_i \in \mathbb{R}^{d_i \times d}$, $d_i \in \mathbb{N}_+$, $\xi_i \sim \texttt{Gsn}(0, \Sigma_i)$ and $\Sigma_i \in \mathbb{R}^{d_i \times d_i}$ $(i = 1,\ldots,K)$. In particular L_i's are allowed to be rank-deficient, so a setting—frequently encountered in practice—of latent (i.e. unobserved) coordinates is just a special case of (3.13). Additionally, the authors simplify and automate a number of computational components of the main algorithm, rendering the overall procedure substantially easier to implement and usually also computationally more efficient. Finally, Bierkens et al. (2018) extended the method to hypoelliptic diffusions of more general kind

than considered in Delyon and Hu (2006) and not only under conditioning on the end-point, but also on (3.13).

The formulation of Schauer et al. (2017) has been applied to Bayesian inference for diffusion processes (van der Meulen and Schauer, 2017a, 2018); however, more recent ones of van der Meulen and Schauer (2017b) and Bierkens et al. (2018) lack some final steps automating computations of the likelihood under the auxiliary measure (defined below) and in their current form cannot be applied to inference (at least not in their full generality). Additionally, under the assumption of hypoellipticity or when the matrices L_i $(i = 1, \ldots, K)$ are rank-deficient, some of the computations increase in complexity with the size of the dataset. This can be partially offset with the blocking technique (see chapter 5); however, even then, undesirable costs may be incurred (see section 6.1). In chapter 6 I show how to finalise the formulations of van der Meulen and Schauer (2017b) and Bierkens et al. (2018) to automate all the remaining routines, completing one part of the picture of guided proposals as a flexible and efficient method of simulating conditioned diffusions, which once fully implemented requires little to no coding from the user. I also apply such newly formulated algorithm to a problem of Bayesian inference for diffusion processes. Finally, in chapter 7 I extend the method to novel observational regimes.

3.3.1 Doob's h-transform

Guided proposals are based on a fundamental result from stochastic analysis: Doob's h-transform (Rogers and Williams, 2000b, Chapter IV.39). This technique, when applied to diffusion processes, allows to rigorously define a conditioned diffusion as an unconditioned one by modifying the drift of the original stochastic differential equation.

In full generality, let X be some continuous time Markov process on a filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P})$ with the transition kernel $p(t, t+s, x, y)\,\mathrm{d}y := \mathbb{P}(X_{t+s} \in \mathrm{d}y | X_t = x)$ (which at times I will also denote with $\mathrm{d}\mathbb{P}(X_{t+s} = y | X_t = x)$), and let $h(t, x)$ define a *space-time harmonic* function for X, i.e. a function equipped

with the following property:

$$h(t,x) = \int_{\mathscr{X}} p(t,t+s,x,u)h(t+s,u)\,\mathrm{d}u = \mathbb{E}[h(t+s,X_{t+s})|X_t = x],$$
$$\forall t \in [0, T-s], s \in [0,T], x \in \mathscr{X}.$$

In other words, h needs to be such that $Z := \{h(t,X_t);\, t \in [0,T]\}$ is a continuous local martingale. Then, it is possible to define a new measure $\mathbb{Q}$ on $(\Omega, \mathscr{F})$ via

$$\left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_t := (h(0,x_0))^{-1} Z_t = \frac{h(t,X_t)}{h(0,x_0)}. \tag{3.14}$$

Under this new measure $\mathbb{Q}$, the transition kernel of the process X is given by:

$$q(t,t+s,x,y) = \frac{h(t+s,y)}{h(t,x)} p(t,t+s,x,y). \tag{3.15}$$

Proof. Notice:

$$q(t,t+s,x,y)\,\mathrm{d}y = \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_{\{\mathrm{d}y\}}(X_{t+s})|X_t = x\right] = \mathbb{E}_{\mathbb{P}}\left[\left.\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right|_t \left.\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right|_{t+s} \mathbb{1}_{\{\mathrm{d}y\}}(X_{t+s}) \,\middle|\, X_t = x\right],$$

from which the result follows. □

Suppose that $\mathcal{Z}$ is such that $h(t,x) := \mathrm{d}\mathbb{P}(\mathcal{Z}|X_t = x)$ is space-time harmonic. Then, by the Markov property, eq. (3.15) takes a form

$$q(t,t+s,x,y) = \frac{\mathrm{d}\mathbb{P}(\mathcal{Z}|X_{t+s} = y)}{\mathrm{d}\mathbb{P}(\mathcal{Z}|X_t = x)} \mathrm{d}\mathbb{P}(X_{t+s} = y|X_t = x) = \mathrm{d}\mathbb{P}(X_{t+s} = y|X_t = x, \mathcal{Z}).$$

This means that under the measure $\mathbb{Q}$, process X behaves exactly the same as the conditioned process $X|\mathcal{Z}$ does under the measure $\mathbb{P}$, or in other words, that $\mathbb{Q}$ coincides with the conditioned law $\mathbb{P}(\cdot|\mathcal{Z})$.

In the context of diffusion processes, X is a diffusion solving the stochastic differential equation (1.1) and $h(t,x) := \mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_t = x)$ for a random variable (or an event) $\mathcal{Z}$ of interest. Provided that such choice of $\mathcal{Z}$ does indeed result in a space-time harmonic h, the new diffusion law $\mathbb{Q}$ defined by (3.14) coincides with the conditioned diffusion law $\mathbb{P}_b(\cdot|\mathcal{Z})$. Notice that space-time harmonicity of h is equivalent to $(\partial_t + \mathscr{L})h = 0$, for $\mathscr{L}$ denoting an infinitesimal generator of X (given in eq. (2.7)). Indeed, by Itō's Lemma:

$$\mathrm{d}h(t,X_t) = [\nabla_x h(t,X_t)]^T \sigma(t,X_t)\,\mathrm{d}W_t + (\partial_t + \mathscr{L})h(t,X_t)\,\mathrm{d}t,$$

and the rightmost term must vanish if $h(t,X_t)$ is to define a local martingale. Additionally, notice that the generator of X under the new law $\mathbb{Q}$ is given by:

$$\mathscr{G} = (\partial_t + \mathscr{L})f + (\Gamma\nabla_x \log h)^T \nabla_x f. \tag{3.16}$$

Proof.

$$\begin{aligned}
[\mathscr{G}f](t,x) &:= \lim_{s\downarrow 0} \frac{\mathbb{E}_{\mathbb{Q}}[f(t+s,X_{t+s})|X_t = x] - f(t,x)}{s} \\
&= \lim_{s\downarrow 0} \frac{\mathbb{E}_{\mathbb{P}}[f(t+s,X_{t+s})\frac{h(t+s,X_{t+s})}{h(t,x)}|X_t = x] - f(t,x)}{s} \\
&= \frac{1}{h(t,x)} \lim_{s\downarrow 0} \frac{\mathbb{E}_{\mathbb{P}}[[fh](t+s,X_{t+s})|X_t = x] - [fh](t,x)}{s} \\
&= \frac{1}{h(t,x)}[(\partial_t + \mathscr{L})(fh)](t,x) \\
&= \left[\frac{f}{h}\cdot(\partial_t + \mathscr{L})h + (\partial_t + \mathscr{L})f + \frac{1}{h}(\Gamma\nabla_x h)^T\nabla_x f\right](t,x) \\
&= \left[(\partial_t + \mathscr{L})f + (\Gamma\nabla_x \log h)^T \nabla_x f\right](t,x)
\end{aligned}$$

□

It is now straightforward to write down the stochastic differential equation that the process X solves under the new law $\mathbb{Q}$ (which, recall, is equal to $\mathbb{P}_b(\cdot|\mathcal{Z})$)

$$\mathrm{d}X_t = [b(X_t) + \Gamma(X_t)\nabla_x \log h(t,X_t)]\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0,T]. \tag{3.17}$$

Equation (3.17) takes a form of an unconditioned stochastic differential equation and yet, paths which solve it follow the conditioned law $\mathbb{P}_b(\cdot|\mathcal{Z})$. This seemingly magical property is a result of a presence of the extra *pulling term* $\Gamma(X_t)\nabla_x \log h(t,X_t)$. Indeed, it pulls the diffusion paths in just the right way, so that the path is consistent with $\mathcal{Z}$ and the resulting law is given exactly by $\mathbb{P}_b(\cdot|\mathcal{Z})$. The following example is instructive in understanding how the term $\Gamma\nabla \log h$ achieves this feat.

Example 3.3.1. Let $\mathcal{Z} := X_T$ and set $h(t,x) := \mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_t = x)$. Then h is space-time harmonic.

Proof. In this case $\mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_t = x) = p(t,T,x,x_T)$, and thus the Chapman-Kolmogorov equations yield

$$\begin{aligned}\int_{\mathscr{X}} p(t,t+s,x,u)h(t+s,u)\,\mathrm{d}u &= \int_{\mathscr{X}} p(t,t+s,x,u)p(t+s,T,u,x_T)\,\mathrm{d}u \\ &= p(t,T,x,x_T) \\ &= h(t,x).\end{aligned}$$

□

This shows that eq. (3.17) can be used to simulate diffusion bridges. Now, restrict attention to a special case of eq. (1.1): a d-dimensional Brownian motion. In this case $b = 0$ and $\sigma = I_d$ and h is simply a Gaussian probability density function (pdf). Elementary calculations yield

$$\Gamma(x)\nabla_x \log h(t,x) = \frac{1}{T-t}(x_T - x), \tag{3.18}$$

and thus the SDE for Brownian motion conditioned on an end-point takes the following form:

$$\mathrm{d}X_t = \frac{1}{T-t}(x_T - X_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0,T].$$

Notice how the force of the pulling term $\frac{1}{T-t}(x_T - x)$ is modulated via $(T-t)^{-1}$. Diffusion X is pulled towards a point x_T—where the process should end-up—on the entirety of the interval $[0,T]$; however, the force increases as $t \uparrow T$, to the point of explosion. Close to the terminal point T any fluctuations from $\mathrm{d}W_t$ term are being dwarfed by the pulling term, so that X must ultimately end-up in x_T.

Simulation of paths from the conditioned diffusion law $\mathbb{P}_b(\cdot|\mathcal{Z})$ can thus be achieved by forward-simulating unconditioned diffusion paths from eq. (3.17). Unfortunately, this conclusion has one major oversight. With the exception of the simplest of diffusions, $h(t,x) := \mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_t = x)$ is intractable.

3.3.2 Choice of proposals

The idea of guided proposals is initiated with the following inquiry. Is it possible to substitute an intractable term $\Gamma(x)\nabla_x \log h(t,x) := \Gamma(x)\nabla_x \log \mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_t = x)$ with some other, tractable one, good enough so that it mimics the role of $\Gamma\nabla \log h$

of a pulling term, forcing the diffusion to be consistent with $\mathcal{Z}$? Is it possible to devise an even less intrusive approximation and find a suitable term $\widetilde{h}$ to use as a substitution for h? As Clark (1990) and Delyon and Hu (2006) show for $\mathcal{Z} := X_T$, the answer to the former question is affirmative. Schauer et al. (2017) additionally prove the same conclusion for the latter question. Moreover, they show that the range of choices for a tractable term $\widetilde{h}$ is reasonably broad. The proofs of Schauer et al. (2017), though beyond the scope of this thesis, are general enough to translate to other forms of $\mathcal{Z}$, as used for instance in van der Meulen and Schauer (2018, Theorem 3.3) and van der Meulen and Schauer (2017b) to apply to $\mathcal{Z}$ of the form in eq. (3.13). I show in chapter 7 how to apply the idea of guided proposals to $\mathcal{Z}$ concerning first passage time events.

The general recipe of Schauer et al. (2017) is to find another, auxiliary process $\widetilde{X}$, inducing law $\widetilde{\mathbb{P}}$, and substitute h in eq. (3.17) for $\widetilde{h}(t,x) := \mathrm{d}\widetilde{\mathbb{P}}(\mathcal{Z}|\widetilde{X}_t = x)$. This yields the following stochastic differential equation

$$\begin{aligned} \mathrm{d}X_t = b^\circ(t,X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0,T], \\ \text{where } b^\circ(t,x) := b(x) + \Gamma(x)\nabla_x \log \widetilde{h}(t,x). \end{aligned} \tag{3.19}$$

If the auxiliary law $\widetilde{\mathbb{P}}$ is chosen carefully enough, then paths $X \sim \mathbb{P}_{b^\circ}$ simulated from the unconditioned law induced by (3.19) do indeed end-up in $\mathcal{Z}$. Additionally, the laws $\mathbb{P}_b(\cdot|\mathcal{Z})$ and $\mathbb{P}_{b^\circ}$ are absolutely continuous with respect to one another and it is possible to derive the Radon-Nikodým derivative between the two.

Schauer et al. (2017) give a set of conditions that the auxiliary law $\widetilde{\mathbb{P}}$ needs to satisfy and show in Schauer et al. (2017, Theorem 2) that the family of linear diffusions:

$$\begin{aligned} \mathrm{d}\widetilde{X}_t = \widetilde{b}(t,\widetilde{X}_t)\,\mathrm{d}t + \widetilde{\sigma}_t\,\mathrm{d}W_t, \quad \widetilde{X}_0 = x_0, \quad t \in [0,T], \\ \text{where } \widetilde{b}(t,x) := \widetilde{B}_t x + \widetilde{\beta}_t, \end{aligned} \tag{3.20}$$

satisfy those conditions when $\mathcal{Z} = X_T$, provided the extra *matching* condition C7 holds

Condition C7. *Volatility coefficients of the auxiliary and the target diffusions match at the end-point:* $\widetilde{\sigma}_T = \sigma(X_T)$.

They additionally assume that a set of regularity conditions holds, which, *inter alia*, simplify proofs of the existence of strong solutions to eq. (3.20) (and include Lipschitz continuity, linear growth and boundedness of σ and b); however, working with diffusions that fall outside of this family is possible, see Schauer et al. (2017) for details. van der Meulen and Schauer (2018) directly extend the results of Schauer et al. (2017) to the conditioning of the form in eq. (3.13). Both of these results rest on the additional assumption of uniformly ellipticity (A3)

This has been further extended in Bierkens et al. (2018) to some hypoelliptic diffusions (i.e. when A4 holds but not A3), at an expense of imposing additional assumptions:

Assumption A14. *(Bierkens et al., 2018, Assumption 2.7) There exists an invertible $m \times m$ diagonal matrix-valued function $S(t)$ (where m comes from the $m \times d$ observational operators L_i), which is measurable on $[0,T]$, a $t_0 < T$, $\gamma \in (0,1]$ and positive constants $\underline{c}, \overline{c}, c_1, c_2$ and c_3 such that for all $t \in [t_0, T)$*

$$\begin{aligned}
\underline{c}(T-t)^{-1} \leq \lambda_{\min}(M_S(t)) \leq \lambda_{\max}(M_S(t)) \leq \overline{c}(T-t)^{-1}, \\
\left\| [L_S(\widetilde{b} - b)](t,x) \right\| \leq c_1, \\
\mathtt{tr}\left([L_S \Gamma L_S^T](t,x) \right) \leq c_2, \\
\left\| [L_S(\widetilde{\Gamma} - \Gamma) L_S^T](t,x) \right\| \leq c_3 (T-t)^{\gamma},
\end{aligned}$$

where $L_S(t) := [S\widetilde{L}](t)$, $M_S(t) := [S^{-1}\widetilde{M}S^{-1}](t)$, and the pair $\widetilde{L}(t)$, $\widetilde{M}(t)$ are defined in eq. (6.4) *and eq.* (6.7) *respectively.*

$\lambda_{\min}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote respectively the smallest and the largest eigenvalue. The second inequality above is perhaps the most interesting and essentially calls for the corresponding *matching condition* C8 on the drift function $\widetilde{b}$:

Condition C8. *Drift coefficients of the auxiliary and the target diffusions match at the end-point: $\widetilde{b}_T = b(X_T)$.*

Let me remark that other, technical conditions that are imposed on the drift or volatility coefficients at various points of Schauer et al. (2017) and Bierkens et al. (2018)—and which I briefly mentioned above—appear mainly as technical tools that facilitate proofs and often bear little consequence on the simulation procedure

(which, unlike exact rejection sampling on a path space, at its core does not stray away from admitting approximation errors in favour of extending its applicability). For instance, to satisfy boundedness assumption, the original SDE (1.1) from which paths are to be simulated, can be changed, so that the drift b and the volatility σ are cropped at some arbitrarily large hypercube. The boundaries of such hypercube can be made arbitrarily large, so that during simulations at no point do the drift or volatility coefficients need to be evaluated outside of them. Notice that this is different from the relationship of A10 to rejection sampling on a path space say, as in the latter case the value of a lower bound l_* of φ directly influences the simulation procedure.

Example 3.3.2. Suppose that $\mathcal{Z} := X_T$ and σ is a constant matrix. Take $\widetilde{\mathbb{P}}$ to be the law of a scaled, d-dimensional Brownian motion σW_t. Then

$$\Gamma(x)\nabla_x \log \widetilde{h}(t,x) = \frac{1}{T-t}(x_T - x).$$

In particular, eq. (3.19) takes the following form

$$\mathrm{d}X_t = \left[b(X_t) + \frac{1}{T-t}(x_T - X_t) \right] \mathrm{d}t + \sigma \, \mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T].$$

Example 3.3.2 is quite unusual in that Delyon and Hu (2006) bridges can be seen as a special case of Schauer et al. (2017) bridges. Delyon and Hu (2006) bridges are in fact based on a different association and can only coincide with those of Schauer et al. (2017) if the volatility coefficient of the target diffusion (1.1) is state-independent. Conversely, the class of Schauer et al. (2017) bridges **is not** a subclass of Delyon and Hu (2006) bridges and leads to different types of proposal diffusions. In fact this is what allowed Schauer et al. (2017), van der Meulen and Schauer (2017b) and Bierkens et al. (2018) to extend the idea of guided proposals to more general settings than presented in Delyon and Hu (2006). To be more explicit, Delyon and Hu (2006) bridges are defined to have either of the two forms below:

$$\mathrm{d}X_t = \frac{1}{T-t}(x_T - X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T],$$
$$\mathrm{d}X_t = \left[b(X_t) + \frac{1}{T-t}(x_T - X_t) \right] \mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T].$$

Consequently it is the entire terms $(b + \Gamma\nabla \log h)$ or $(\Gamma\nabla \log h)$ from eq. (3.17) that are being substituted with the corresponding term obtained for a scaled, d-dimensional Brownian motion. The substitution of Schauer et al. (2017) is less invasive and can be additionally moulded through more appropriate choices of the auxiliary diffusion (3.20), so as to better match $\widetilde{h}$ and h. Figure 3.2 illustrates how a choice of the auxiliary process can influence the law of the proposal in eq. (3.19).

3.3.3 Correcting discrepancies in law

Naturally, exchanging h for a tractable $\widetilde{h}$ renders the law of the proposals $X \sim \mathbb{P}_{b^\circ}$ different from the law of the target $\mathbb{P}_b(\cdot|\mathcal{Z})$. This discrepancy can however be quantified with the likelihood. Absolute continuity of the proposal and the target laws restricted to any $\mathscr{F}_t$ with $t < T$ follows from Girsanov theorem. Schauer et al. (2017, Proposition 1) formulate the following statement:

Proposition 3.3.1. (Schauer et al., 2017, Proposition 1) Under regularity conditions on $\widetilde{h}$ (satisfied by linear diffusions), the Radon-Nikodým derivative between the two laws $\mathbb{P}_b|_t(\cdot|\mathcal{Z})$ and $\mathbb{P}_{b^\circ}|_t$ is given by

$$\frac{\mathrm{d}\mathbb{P}_b|_t(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_{b^\circ}|_t(X)} = \frac{\widetilde{h}(0,x_0)}{h(0,x_0)}\frac{h(t,X_t)}{\widetilde{h}(t,X_t)}\exp\left\{\int_0^t G(s,X_s)\,\mathrm{d}s\right\}, \quad t < T,$$

$$\text{where} \quad G(s,x) := [(b-\widetilde{b})^T\widetilde{r}](s,x) - \frac{1}{2}[(\Gamma-\widetilde{\Gamma}) : (\widetilde{H}-\widetilde{r}\widetilde{r}^T)](s,x),$$

$$\text{and} \quad \widetilde{r} := \nabla_x \log\widetilde{h}, \quad \widetilde{H} := -D^2\log\widetilde{h},$$

where $D^2_{i,j}f(t,x) := \partial^2 f(t,x)/(\partial x_i \partial x_j)$.

Proof. (Schauer et al., 2017) Before considering the pair $(\mathbb{P}_b|_t(\cdot|\mathcal{Z}), \mathbb{P}_{b^\circ}|_t)$, consider $(\mathbb{P}_b|_t, \mathbb{P}_{b^\circ}|_t)$ instead. Girsanov theorem asserts that the two laws are equivalent and the Radon-Nikodým $\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_{b^\circ}}|_t(X)$ between them is given by eq. (2.2), with u satisfying $\sigma u = b - b^\circ$. Since by definition (3.19): $b^\circ := b + \sigma\sigma^T\widetilde{r}$, it follows that $u = -\sigma^T\widetilde{r}$ and thus:

$$\left.\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_{b^\circ}}\right|_t(X) = \exp\left\{-\int_0^t [\sigma^T\widetilde{r}](s,X_s)\,\mathrm{d}W_s - \frac{1}{2}\int_0^t [\widetilde{r}^T\Gamma\widetilde{r}](s,X_s)\,\mathrm{d}s\right\}. \tag{3.21}$$

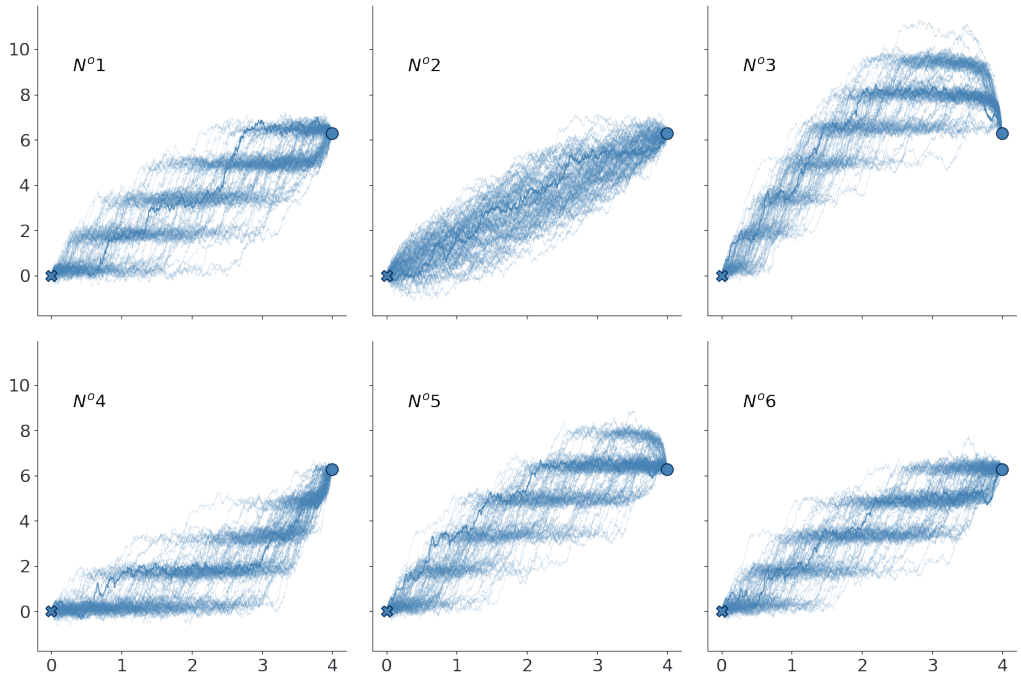


Figure 3.2: Illustration of an impact that a choice of the auxiliary law has on the quality of the proposals. $N°1$: paths of diffusion bridges joining 0 and 2π over the interval $[0,4]$ with unconditioned diffusion law $\mathbb{P}_b$ given by a the sine diffusion: $\mathrm{d}X_t = (4-4\sin(4X_t))\mathrm{d}t + \mathrm{d}W_t$ (the paths for this plot were simulated using an MCMC algorithm on a path space taking sufficiently large number of steps). This is a very difficult, one-dimensional diffusion model exhibiting multi-modality of transition densities. In addition, the end-point is chosen in a construed way—it is unlikely to observe it under the unconditioned diffusion law. $N°2$: paths of 0–2π Brownian bridges. No aspect of the multi-modality is captured, making $\mathbb{P}_0(\cdot|\mathcal{Z})$ decidedly inadequate for sampling from $\mathbb{P}_b(\cdot|\mathcal{Z})$. $N°3$: Delyon and Hu (2006) bridges (which for a unit volatility coefficient coincide with the guided proposals of Schauer et al. (2017) with the choice of Brownian motion for $\widetilde{\mathbb{P}}$). The multi-modality is well-captured; however, since the end-point is unlikely to be observed under the unconditioned law $\mathbb{P}_b$, in the initial stages of the interval $[0,T]$, $\Gamma\nabla\log\widetilde{h}$ does not penalise the positive drift strongly enough, leading to banana-shaped paths that are not representative of true paths under $\mathbb{P}_b(\cdot|\mathcal{Z})$. $N°4$: guided proposals with $\widetilde{\mathbb{P}}$ chosen to be a drifted Brownian motion W_t+4t. $\widetilde{\mathbb{P}}$ is chosen so that the positive drift of SINE diffusion is accounted for, leading to a substantial shift in the mass of the law $\mathbb{P}_{b^\circ}$. Nonetheless, the shift is too strong, because the end-point takes an unusually small value. $N°5$: guided proposals with $\widetilde{\mathbb{P}}$ chosen to be a drifted Brownian motion $W_t+0.5\pi t$. Instead of taking a drift of an unconditioned process, a slope of a line joining start-point and end-point is chosen. Substantially better results are obtained, albeit the end-point is approached from the top and the bottom alike, in contrast to what is observed under $\mathbb{P}_b(\cdot|\mathcal{Z})$. $N°6$: guided proposals with $\widetilde{\mathbb{P}}$ chosen to be a linear diffusion solving: $\mathrm{d}\widetilde{X}_t = (\frac{t}{5T}\widetilde{X}_t + 0.5\pi)\mathrm{d}t + \mathrm{d}W_t$. An additional term $\frac{t}{5T}x$, which increases in magnitude towards the end-point aims at penalising paths under $\mathbb{P}_{b^\circ}$ for approaching the end-point from above. The law of the resulting bridges closely match the target law $\mathbb{P}(\cdot|\mathcal{Z})$.

Denote by $\mathscr{L}$, $\mathscr{L}^\circ$ and $\widetilde{\mathscr{L}}$ the infinitesimal generators under the laws $\mathbb{P}_b$, $\mathbb{P}_{b^\circ}$ and $\mathbb{P}_{\tilde{b}}$ respectively and notice that $\mathscr{L}^\circ f := \mathscr{L} + (\Gamma\tilde{r})^T \nabla f$ (as shown in eq. (3.16)). Now, by Itō's Lemma:

$$\begin{aligned}\mathrm{d}\Big(\log\tilde{h}(t,X_t)\Big) &= [(\partial_t + \mathscr{L}^\circ)\log\tilde{h}](t,X_t)\,\mathrm{d}t + [(\nabla_x \log\tilde{h})^T\sigma](t,X_t)\,\mathrm{d}W_t \\ &= [(\partial_t + \mathscr{L})\log\tilde{h} + \tilde{r}^T\Gamma\tilde{r}](t,X_t)\,\mathrm{d}t + [\tilde{r}^T\sigma](t,X_t)\,\mathrm{d}W_t.\end{aligned}$$

Notice additionally, that

$$\begin{aligned}(\partial_t + \widetilde{\mathscr{L}})\log\tilde{h} &= \frac{1}{\tilde{h}}(\partial_t + \widetilde{\mathscr{L}})\tilde{h} - \frac{1}{2\tilde{h}^2}\Gamma : \Big(\nabla\tilde{h}[\nabla\tilde{h}]^T\Big) \\ &= -\frac{1}{2}[\nabla\log\tilde{h}]^T\tilde{\Gamma}[\nabla\log\tilde{h}] \\ &= -\frac{1}{2}\tilde{r}^T\tilde{\Gamma}\tilde{r},\end{aligned}$$

where it was used that $\tilde{h}$ is space-time harmonic under the measure $\mathbb{P}_{\tilde{b}}$, i.e. $(\partial_t + \widetilde{\mathscr{L}})\tilde{h} = 0$. Combining the two equations above yields

$$\mathrm{d}\Big(\log\tilde{h}(t,X_t)\Big) = [(\mathscr{L}-\widetilde{\mathscr{L}})\log\tilde{h} + \frac{1}{2}\tilde{r}^T\Big(\Gamma-\tilde{\Gamma}\Big)\tilde{r} + \frac{1}{2}\tilde{r}^T\Gamma\tilde{r}](t,X_t)\,\mathrm{d}t + [\tilde{r}^T\sigma](t,X_t)\,\mathrm{d}W_t.$$

Comparing this with (3.21), results in

$$\left.\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_{b^\circ}}\right|_t(X) = \frac{\tilde{h}(0,x_0)}{\tilde{h}(t,X_t)}\exp\left\{\int_0^t\left[(\mathscr{L}-\widetilde{\mathscr{L}})\log\tilde{h} + \frac{1}{2}\tilde{r}^T\Big(\Gamma-\tilde{\Gamma}\Big)\tilde{r}\right](s,X_s)\,\mathrm{d}s\right\}. \tag{3.22}$$

Since $\mathbb{P}_b|_t(X|\mathcal{Z})$ can be seen as the law induced by eq. (3.17), it follows from eq. (3.14) that

$$\frac{\mathrm{d}\mathbb{P}_b|_t(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_b|_t(X)} = \frac{h(t,X_t)}{h(0,x_0)}. \tag{3.23}$$

The result follows after combining (3.22), (3.23) and realising:

$$(\mathscr{L}-\widetilde{\mathscr{L}})\log\tilde{h} = (b-\tilde{b})^T\tilde{r} - (\Gamma-\tilde{\Gamma}) : \tilde{H}.$$

□

In the presence of noise on the observations it can be verified by direct calculations (using for instance van der Meulen and Schauer (2017b, §2.2)) that the term $\tilde{r}$ stays Lipschitz on the interval $[0,T]$ and thus Girsanov theorem applies also in the limit $t \uparrow T$. For details, see the proof of van der Meulen and Schauer (2018,

Theorem 3.3). Unfortunately, if no noise is present, then the situation is no longer so simple. For elliptic diffusion bridges Schauer et al. (2017) present a rigorous justification for taking the limit $t \uparrow T$ in a no-noise setting, which is then expanded to general observations in van der Meulen and Schauer (2018, Theorem 3.3) (which includes the possibility of presence of noise only on a subset of the coordinates). These proofs are then expanded upon in Bierkens et al. (2018) to target hypoelliptic diffusions. It is shown that $\lim_{t\uparrow T}\left(\frac{h(t,X_t)}{\widetilde{h}(t,X_t)}\right) \to 1$ and the limits inside the integral are expanded from 0–t to 0–T. Consequently, the following result is presented:

Theorem 3.3.1. (Schauer et al., 2017, Theorem 1& 2), (van der Meulen and Schauer, 2018, Theorem 3.3), (Bierkens et al., 2018, Theorem 2.14) If the auxiliary law $\widetilde{\mathbb{P}}$ is induced by a linear diffusion of the form in eq. (3.20), then under uniform-ellipticity A3, (and in case of no-noise, also a matching condition C7):

$$\frac{\mathrm{d}\mathbb{P}_b(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_{b^\circ}(X)} = \frac{\widetilde{h}(0,x_0)}{h(0,x_0)} \exp\left\{\int_0^T G(s,X_s)\,\mathrm{d}s\right\}, \tag{3.24}$$

where G, $\widetilde{r}$ and $\widetilde{H}$ are defined in proposition 3.3.1. Additionally, under hypoellipticity A4 (and in case of no-noise, a matching condition C7) and assumption A14, eq. (3.24) holds as well.

Equation (3.24) can be used to design a Markov chain Monte Carlo algorithm that targets a diffusion law $\mathbb{P}_b(\cdot|\mathcal{Z})$ using proposals from $\mathbb{P}_{b^\circ}$. Notice that for a fixed $\mathcal{Z}$, $\frac{\widetilde{h}(0,x_0)}{h(0,x_0)}$ is just a constant and thus algorithm 3.5 below can be used to sample paths under $\mathbb{P}_b(\cdot|\mathcal{Z})$.

Algorithm 3.5 Independent Metropolis-Hastings sampling using guided proposals

1: Draw $X^{(0)} \sim \mathbb{P}_{b^\circ}$ ▷ This has unconditioned form
2: **for** $n = 0,\ldots,N-1$ **do**
3: Draw $X^\circ \sim \mathbb{P}_{b^\circ}$ ▷ This has unconditioned form
4: Draw $E \sim \mathrm{Exp}(1)$
5: **if** $E \geq \int_0^T G(s,X_s^{(n)})\,\mathrm{d}s - \int_0^T G(s,X_s^\circ)\,\mathrm{d}s$ **then**
6: Set $X^{(n+1)} \leftarrow X^\circ$
7: **else**
8: Set $X^{(n+1)} \leftarrow X^{(n)}$
9: **return** $\{X^{(n)}; n = 0,\ldots,N\}$ ▷ Markov chain with the invariant density $\mathrm{d}\mathbb{P}_b(\cdot|\mathcal{Z})$

All path-sampling steps need to forward-simulate paths from unconditioned stochastic differential equations and thus can be completed with the use of very efficient sampling methods based on stochastic Taylor expansions (Kloeden and Platen, 2013).

3.3.4 Discussion

It is intuitively clear why guided proposals can be expected to approximate the target law well, regardless of whether the observations are dense or sparse. The argument is summarised graphically in fig. 3.3. If T is small, then the discussion from section 2.3.4 makes it clear that the approximate $\widetilde{h}$ closely matches h and therefore the two SDEs (3.17) and (3.19) differ only marginally. On the other hand, if T is large (in a sense $T \to \infty$), then the term $\Gamma\nabla_x \log h$ is small on the initial part of the interval and increases as t approaches T (one can become convinced of this fact after examining the form of $\Gamma\nabla_x \log h$ in a simple example 3.3.2). This is a reciprocal relation to how $\widetilde{h}$ matches h. The largest discrepancies between the latter pair are on the initial parts (or, depending on the example, a mid-section) of the interval $[0, T]$ and the differences between the two shrink as $t \uparrow T$. Consequently, when the inadequacy of approximating h with $\widetilde{h}$ is at its worst, the term $\Gamma\nabla_x \log h$ is also at its smallest, and thus the influence of approximation error is conveniently of little relevance for $t \ll T$. Conversely, $\Gamma\nabla \log h$ starts to increase in size only as $t \uparrow T$, but for t close to T, $\widetilde{h}$ closely resembles h and thus the approximation error is once again being conveniently self-regulated.

Additionally, guided proposals of Schauer et al. (2017) hold a major advantage over those of Delyon and Hu (2006)—it is possible to further improve the fidelity with which $\widetilde{h}$ approximates h by picking a better auxiliary process from the family of linear diffusions in eq. (3.20). This choice is arbitrary, but there are certain rules of thumb expected to perform particularly well across a vast range of use cases (van der Meulen and Schauer, 2017a). From the discussion above, it is clear that the SDE (3.19) of a proposal process induces the largest errors on a range leading up to T, i.e. on $[T-\epsilon, T)$, where ϵ is—vaguely speaking—*moderately sized*. Consequently, using an auxiliary process $\widetilde{X}$ that contains a term that increases in its magnitude as $t \uparrow T$ and comes from linearising the target diffusion at its end-point,

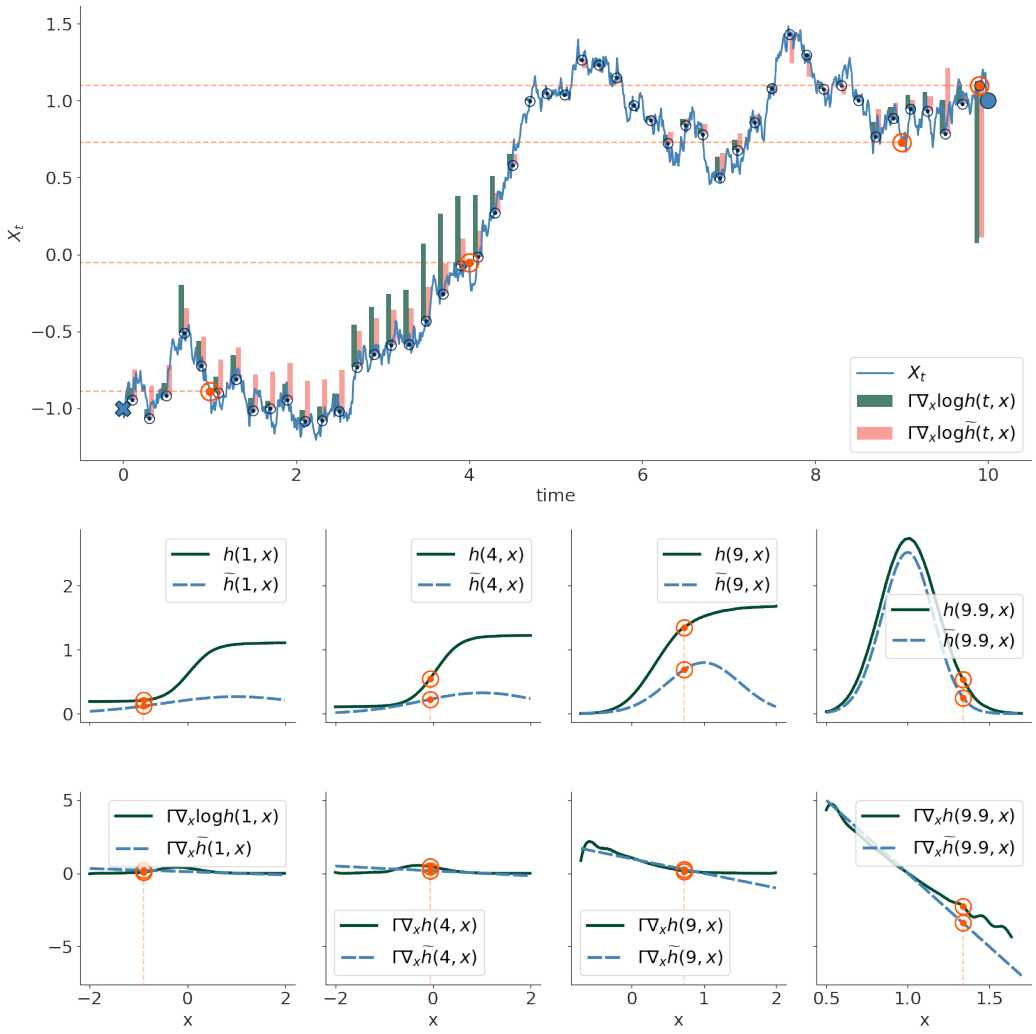


Figure 3.3: Qualitative understanding of the error induced by the substitution $h \to \widetilde{h}$, when $\mathbb{P}_b$ is taken to be the law of double-well potential in eq. (2.1), with $(\rho, \mu, \sigma) = (1, 1, 0.5)$. Top plot gives a single realisation of a path drawn under the proposal measure $\mathbb{P}_{b^\circ}$, where $\widetilde{\mathbb{P}}$ is taken to be the law of a scaled Brownian motion σW. Every 0.2 time units a magnitude of $\Gamma\nabla \log h$ and $\Gamma\nabla \log \widetilde{h}$ at a given point is plotted, illustrating the difference between a force with which guided proposal is pulled ($\Gamma\nabla \log \widetilde{h}$) and a force with which it should be pulled ($\Gamma\nabla \log h$). In the middle row $x \to h(t,x)$ is compared against $x \to \widetilde{h}(t,x)$ at 4 different time points $t \in \{1, 4, 9, 9.9\}$. The value taken by a sampled trajectory X_t at a corresponding time point is marked with an orange circle. In the bottom row analogous plots of $x \to \Gamma(x)\nabla_x \log h(t,x)$ and $x \to \Gamma(x)\nabla \log_x \widetilde{h}(t,x)$ are given.

through say, Taylor expansions of its drift and volatility coefficient, often leads to good performance improvements. Another possibility is to use some of the ideas from Whitaker et al. (2017) (discussed further in section 3.4) to find fitting

candidates from the family of linear diffusions in eq. (3.20) for approximating the target (1.1). The ideas of Whitaker et al. (2017) are based on solving the system of ordinary differential equations that result from removing the stochastic term from eq. (1.1):

$$\mathrm{d}x(t) = b(x(t))\,\mathrm{d}t, \quad x(0) = x_0. \tag{3.25}$$

The simplest way to proceed is to use a solution $x(t)$ of the ODE above to define $\widetilde{\beta}_t := b(x(t))$, $\widetilde{B}_t := 0$ and $\widetilde{\sigma}_t := \sigma(X_T)$ in (3.20) (provided X_T is known from the conditioning set). A higher order approximation would instead use $\widetilde{\beta}_t := b(x(t)) - V(x(t))x(t)$, $\widetilde{B}_t := V(x(t))$ and $\widetilde{\sigma}_t := \sigma(X_T)$ with $V(x) := \nabla_x b(x)$. The impact that the choice of the auxiliary process $\widetilde{X}$ may have on the quality of the proposals is illustrated in figure fig. 3.2.

There are two main components to any MCMC algorithm on a path-space that together determine the overall efficiency of this procedure. The first one is the accuracy with which proposals approximate the target law—guided proposals have been shown to be particularly strong in this respect. The second one is just as important—it is the computational cost associated with each draw of a proposal path. For guided proposals this amounts to computing $\widetilde{r}$ and $\widetilde{H}$ on a time-grid $(t_0, \ldots, t_M)$, sampling a diffusion path $X \sim \mathbb{P}_{b^\circ}$ using, say, the Euler-Maruyama scheme and then, computing the likelihood (3.24) via Riemann sums. The latter two steps are very simple and efficient, and their cost scales linearly in T. If the goal of sampling is to obtain an empirical distribution over a path space, it can be argued that the cost of computing $\widetilde{r}$ and $\widetilde{H}$ is not as important, as the pair needs to be computed only once—then, once computed, an unbounded number of samples can be obtained. Nonetheless, in many applications and particularly the one central to this thesis—of Bayesian inference for diffusion processes—the pair $(\widetilde{r}, \widetilde{H})$ needs to be re-computed on a time grid $(t_0, \ldots, t_M)$ for each new draw of a proposal path (see chapter 4 for details). Consequently, having fast and accurate routines for deriving $(\widetilde{r}, \widetilde{H})$ is crucial.

Derivations of $(\widetilde{r}, \widetilde{H})$, as presented in Schauer et al. (2017) require repeated, expensive computations of matrix exponentials. This has been improved upon in some settings by van der Meulen and Schauer (2017b), where the authors characterise $(\widetilde{r}, \widetilde{H})$ as appropriately transformed solutions to a system of ordinary differ-

ential equations (so that only simple and efficient matrix multiplications, additions and inversions are ever used). Whenever applicable, this modification drastically decreases the computational cost of the algorithm, and essentially puts the computation of $(\widetilde{r}, \widetilde{H})$ on the same scale as the subsequent Euler-Maruyama pass as well as the final step computing the likelihood. Unfortunately, in some settings, the solutions to those ODEs increase in their dimension with the size of the dataset, creating unnecessary losses in efficiency. Furthermore, for Bayesian inference for diffusion processes, $\widetilde{h}(0, x_0)$ needs to be evaluated alongside $(\widetilde{r}, \widetilde{H})$ and van der Meulen and Schauer (2017b) do not give any methods for computing it efficiently. In chapter 6 I show how to solve both of those issues, so that guided proposals can always be implemented in an efficient manner (including in the setting of Bayesian inference for diffusion processes).

3.4 Review of alternative methods

Statistics literature is littered with various methods for simulating conditioned diffusion processes, though vast majority of them are suitable solely to the case of diffusion bridges or conditioning on the noisy and partial observations of the process as in eq. (3.13). In this section I give a brief review of these alternative algorithms.

Naturally, the simplest way of simulating conditioned diffusions is to simply define a rejection sampler that draws paths from an unconditioned law and accepts only those paths for which $\mathcal{Z} = \mathbb{z}$ occurs, for a desired $\mathbb{z}$. Of course, for a number of interesting cases of $\mathcal{Z}$—such as conditioning on the value of the end-point—the conditioned-on event happens with probability zero, rendering such sampling strategy infeasible. For diffusion bridges a simple relaxation of the acceptance condition to $\mathbb{1}_{(x_T-\epsilon, x_T+\epsilon)}(X_T)$ for some small enough $\epsilon > 0$ can lead to an admissible (albeit highly inefficient) approximation. Alternatively, as proposed by Pedersen (1995), the unconditioned diffusion can be simulated only up to time $(T - \epsilon)$ and such sample can then be weighted in an importance sampling setting.

Due to their simplicity, modified diffusion bridges introduced by Durham and Gallant (2002) (and subsequently extended to observations with errors in Golightly and Wilkinson (2008)) constitute a popular choice, often made by the practitioners. At its conception, the algorithm has been derived by approximating the joint

distribution of $X_{t_{i+1}}, X_T | X_{t_i}$, for each $i = 1, \ldots, K$ by freezing the coefficients of an SDE (1.1) on $[t_i, T]$, $(i = 0, \ldots, K)$ (resulting in Gaussian laws), and using standard results on the conditional distributions of multivariate Gaussian random variables to subsequently derive the distributions of $X_{t_{i+1}} | (X_T, X_{t_i}) = (x_T, x)$:

$$\left[X_{t_{i+1}} | (X_T, X_{t_i}) = (x_T, x)\right] \sim \texttt{Gsn}\left\{x + \frac{x_T - x}{T - t_i}\Delta, \frac{T - t_{i+1}}{T - t_i}\Gamma(x)\right\}. \tag{3.26}$$

This viewpoint has an advantage of involving only simple mathematics and it extends easily to observations of linearly transformed process X with additive Gaussian noise (see Golightly and Wilkinson (2008)); however, it is not clear from it, whether the approximating bridges come from a law that is even absolutely continuous with respect to the target law $\mathbb{P}_b(\cdot|\mathcal{Z})$. In this respect, presentation of Papaspiliopoulos et al. (2013) is preferable. There, an update equation eq. (3.26) is derived by employing the linear approximation due to Shoji and Ozaki (1998b) and Shoji (1998) to Delyon and Hu (2006) bridges:

$$\mathrm{d}X_t = \frac{1}{T - t}(x_T - X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T].$$

This unmasks a close connection between the modified diffusion bridges and Delyon and Hu (2006) bridges and crucially—at least in the case of exactly observed end-points—allows to inherit the results on the absolute continuity of the target and proposal laws directly from Delyon and Hu (2006).

Lindström (2012) further extends this method by proposing what is heuristically a convex combination of the modified diffusion bridges and an unconditioned dynamics of an original SDE.

Whitaker et al. (2017) propose bridges based on residual processes. First, an ODE in eq. (3.25) is set up by removing the stochastic term from eq. (1.1). The solution to this ODE is then subtracted from the drift of the SDE in eq. (1.1) in hope of removing the strongest non-linear effects. As a result, the following stochastic differential equation is obtained:

$$\mathrm{d}R_t = [b(R_t + x(t)) - b(x(t))]\,\mathrm{d}t + \sigma(R_t + x(t))\,\mathrm{d}W_t, \quad R_0 = 0, \quad t \in [0, T]. \tag{3.27}$$

It is easy to see that $R_t + x(t)$ solves the SDE in eq. (1.1) and thus to simulate $X_t|X_T$ it is enough to simulate $[R_t|R_T = x_T - x(t)]$ instead and return $R_t + x(t)$. Whitaker et al. (2017) propose to use modified diffusion bridges for this last simulation step.

For the cases when $x(t)$ does not capture the dynamics of the underlying process sufficiently well, Whitaker et al. (2017) propose a further improvement. The dynamics of the residual process R in eq. (3.27) are approximated with a linear diffusion $\widehat{R}$, according to the linear noise approximation (Fearnhead et al., 2014). Then, the mean process $\rho_t := \mathbb{E}[\widehat{R}_t|R_0, R_T]$ admits closed form expression and in the same way that $x(t)$ was subtracted from X_t to eliminate the most pronounced non-linear effects, ρ_t is now subtracted from R_t to better centre the process. Modified diffusion bridges are once again employed, this time to draw paths of $\{R_t - \rho_t|R_T\,;\, t \in [0, T]\}$. All of the aforementioned methods based on residual processes extend readily to observations of linearly transformed process X with additive, Gaussian noise, see Whitaker et al. (2017) for details.

An entirely different approach is pursued in the string of papers which characterise the laws of diffusion bridges as invariant laws of certain stochastic partial differential equations (SPDEs) (Stuart et al., 2004; Hairer et al., 2009, 2005, 2007; Beskos et al., 2008). The methodology is applicable to diffusions with constant volatility coefficients, with drifts that consist of a term in a potential form and an additive linear component, i.e. to processes $X \in \mathbb{R}^d$ solving:

$$\mathrm{d}X_u = AX_u\,\mathrm{d}u + -BB^T\nabla_x V(X_u)\,\mathrm{d}u + \Sigma\,\mathrm{d}W_u, \quad X_0 = x_0, \quad X_U = x_U, \quad u \in [0, U], \tag{3.28}$$

where $A, B, \Sigma \in \mathbb{R}^{d\times d}$, some potential $V : \mathbb{R}^d \to \mathbb{R}$ and where I use index u to denote a time variable in order to distinguish it from a time variable of the forthcoming SPDEs.

The idea rests upon Langevin dynamics in infinite dimensional spaces. The details of the methodology are beyond the scope of this thesis; however, I aim to provide an intuition behind the sampler. In finite dimensions, it is well-known that the stationary distribution of a diffusion can be related to the form of its drift coefficient as follows:

Theorem 3.4.1. (Hairer et al., 2009, Theorem 3) Let $L \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that the SDE:

$$\mathrm{d}Z_t = LZ_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{3.29}$$

has a stationary distribution v. Let $\varphi \in \mathcal{C}^2(\mathbb{R}^d;\mathbb{R})$ be a strictly positive probability density with respect to v. Then, the SDE:

$$\mathrm{d}X_t = (LX_t + \nabla \log \varphi(X_t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

has an invariant distribution given by $\varphi\,\mathrm{d}v$.

Theorem 3.4.1 paves a way for defining an approximate sampler from $\varphi\,\mathrm{d}v$—it is enough to simulate X_T for some large enough $T > 0$ via unconditioned diffusion samplers based on stochastic Taylor expansions (Kloeden and Platen, 2013). If chosen T is large enough, the marginal distribution of X_T will be close to that of $\varphi\,\mathrm{d}v$. The idea behind bridge sampling through SPDEs relies on extending theorem 3.4.1 into infinite dimensional spaces, and force $\varphi\,\mathrm{d}v$ to be the distribution of the conditioned diffusion (1.1): $\mathbb{P}_b(\cdot|\mathcal{Z})$.

Rigorous justification for the extension can be found in Hairer et al. (2005) and Hairer et al. (2007). The SDEs are replaced by stochastic evolution equations with values in real Banach space $E = \mathcal{C}([0,1];\mathbb{R}^d)$, continuously embedded into a real separable Hilbert space $\mathcal{H} := L^2([0,1];\mathbb{R}^d)$. Equation (3.29) is replaced with:

$$\mathrm{d}z_t = \mathcal{L}z_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}w_t, \tag{3.30}$$

where $\mathcal{L} := -\mathcal{C}^{-1}$ for a valid covariance operator $\mathcal{C} : \mathcal{H} \to \mathcal{H}$ (given explicitly in Hairer et al. (2005, Eqs. (3.2) & (3.7))) and w is a cylindrical Wiener process on $\mathcal{H}$. The stationary distribution of (3.30) is given by $\mathtt{Gsn}(m,\mathcal{C})$, with $m \in \mathcal{H}$ (given in Hairer et al. (2005, 2007)). Hairer et al. (2007) show that under certain regularity conditions the stationary distribution of:

$$\mathrm{d}x_t = \mathcal{L}(x_t - m)\,\mathrm{d}t + F(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}w_t, \tag{3.31}$$

with an appropriately chosen drift $F(\cdot)$ (given in Hairer et al. (2007, Eq. (5.5)) or Hairer et al. (2009, p. 9)) admits a conditioned diffusion measure $\mathbb{P}_b(\cdot|\mathcal{Z})$ as its stationary distribution. Therefore, the problem of sampling from $\mathbb{P}_b(\cdot|\mathcal{Z})$ reduces

to the problem of finding $\mathscr{L}$ and F, and then simulating x_T for some large enough T. To this end, eq. (3.31) is written in an SPDE form, as follows:

$$\partial_t x(t,u) = \mathscr{L} x(t,u) - \nabla\Psi(x(t,u)) + \sqrt{2}\partial_t w(t,u), \quad \forall\,(t,u) \in (0,\infty)\times(0,U),$$
$$x(t,0) = x_0, \quad x(t,1) = x_1, \quad \forall\, t \in (0,\infty),$$

for an appropriate function $\Psi(\cdot)$ (given in Hairer et al. (2007, Eq. (5.3))) and is then solved up to time T, using standard numerical techniques for unconditioned SPDEs (Stuart et al., 2004).

3.5 Commentary

Out of the three main algorithms discussed in this chapter, the literature on the *exact* simulation of diffusion bridges received perhaps the most attention, resulting in its diverse and plentiful extensions. A partial list include extensions to jump-diffusion bridges (Pollock, 2013; Casella and Roberts, 2011; Pollock et al., 2016), ϵ-strong simulations (where a piece-wise constant *tunnel* (whose width can be refined to an arbitrary degree) constraining the entire trajectory X is simulated jointly with the skeletal points) (Beskos et al., 2012; Pollock et al., 2016), simulation of the Wright-Fisher diffusions (Jenkins and Spano, 2017) and simulation of diffusions with discontinuous drifts (Papaspiliopoulos et al., 2016). Additionally, within the context of the exact rejection sampler on a path space, for unconditioned diffusions Chen and Huang (2013) describe an alternative method of constructing variable Υ (which gives local bounds to the simulated path). A completely different approach—rooted in the theory of rough paths—has been proposed by Blanchet and Zhang (2017); it relaxes assumptions A6 and A7, but suffers from the infinite computational cost.

4 *Bayesian inference for diffusion processes*

One of the reasons why simulation of conditioned diffusions is such an important topic, is because it constitutes a vital component of the modern Bayesian inference methods for diffusion processes. Typically, some phenomenon is modelled through a stochastic differential equation (1.3), whose form I repeat below for convenience:

$$\mathrm{d}X_t = b_\theta(X_t)\,\mathrm{d}t + \sigma_\theta(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T]. \tag{4.1}$$

The drift and volatility coefficient are allowed to depend on a vector of unknown parameters $\theta \in \Theta$. For a given parameter θ, I also use $\mathbb{P}_b^{(\theta)}$ to denote the law induced by eq. (4.1). Perhaps a more telling statement is that (4.1) defines a family of diffusions, indexed by $\theta \in \Theta$, with some members of $\{\mathbb{P}_b^{(\theta)};\, \theta \in \Theta\}$ describing the true dynamics of the underlying process "better" than others. A priori, the beliefs about the distribution of more appropriate values for the parameter θ are captured through a prior distribution $\pi(\theta)$. The observations of the process X—denoted by $\mathscr{D}$—are then gathered in some form. Some examples of $\mathscr{D}$ can be found at the end of Introduction.

The aim is to derive the *posterior* density over the unknown parameters θ for a given prior $\pi(\theta)$ and an observation set $\mathscr{D}$. It is proportional to:

$$\pi(\theta|\mathscr{D}) \propto \pi(\theta)\pi(\mathscr{D}|\theta), \tag{4.2}$$

where $\pi(\mathscr{D}|\theta)$ denotes the likelihood for observing $\mathscr{D}$. The prior is known and set by the practitioner, whereas the likelihood function is typically intractable. For instance, in the case of exact observations, it is given as a product of transition densities:

$$\pi(\{X_{t_i} = x_i;\, i = 1, \ldots, K\}|\theta) := \prod_{i=1}^{K-1} p^{(\theta)}_{t_{i+1}-t_i}(x_i, x_{i+1}),$$

or in the case of partial and noisy observations it is instead given by:

$$\pi(\{L_i X_{t_i} + \xi_i = v_i\}_{i=1}^K|\theta) := \int_{\mathscr{X}} \cdots \int_{\mathscr{X}} p_{t_1-t_0}(x_0, x_1)\,\mathrm{d}F_K^{(\theta)}(v_K - L_K x_K) \cdot \prod_{i=1}^{K-1} \Big[p^{(\theta)}_{t_{i+1}-t_i}(x_i, x_{i+1})\,\mathrm{d}F_i^{(\theta)}(v_i - L_i x_i) \Big]\,\mathrm{d}x_1 \cdots \mathrm{d}x_N,$$

where $\mathrm{d}F_i^{(\theta)}, (i = 1,\ldots,K)$ denote the densities of $\xi_i, (i = 1,\ldots,K)$. In both expressions above (and in general, in the expressions for $\pi(\mathscr{D}|\theta)$) the transition densities appear, and these are typically intractable. The problem of inference for diffusion processes is thus well-known for its difficulty.

4.1 Inference via data-augmentation

There have been many solutions proposed in the statistics literature to the problem of inference for diffusion processes, I discuss some of them at the end of this section. The solution central to this thesis has been shown to be very successful in a broad range of applications (Beskos et al., 2008; Beskos and Stuart, 2009; Golightly and Wilkinson, 2008), is Bayesian—and thus provides not only estimates for θ, but also, at no extra cost, quantification of uncertainty—and essentially reduces the problem of inference to the problem of sampling from the conditioned diffusion law $\mathbb{P}_b^{(\theta)}(\cdot|\mathscr{D})$. The robustness and efficiency of the final algorithm is inherited from the properties of the corresponding conditioned diffusion sampler.

The underpinning idea is that of data augmentation, which, in the setting of inference for diffusions, goes back to Roberts and Stramer (2001). Notice that had the path X been observed exactly and in its entirety, then the likelihood would have been given by the Girsanov formula (2.3) (or (2.2), for non-invertible Γ). Consequently, it turns out to be easier to consider the joint posterior over a path space and a parameter space: $\pi(\theta, X|\mathscr{D})$, and define a Gibbs sampler which targets this joint posterior by alternately updating unknown parameters (by sampling from $\pi(\theta|X,\mathscr{D})$) and imputing a diffusion path (by sampling from $\pi(X|\theta,\mathscr{D})$). The output of such algorithm is a Markov chain $\{(\theta^{(n)}, X^{(n)}); n = 0,\ldots,N\}$, whose invariant density is given by $\pi(\theta, X|\mathscr{D})$. Then, the marginal chain $\{\theta^{(n)}; n = 0,\ldots,N\}$, obtained by removing all samples of the path X, has the desired invariant density $\pi(\theta|\mathscr{D})$. Algorithm 4.1 summarises these steps.

Notice that since algorithm 4.1 is a Markov chain Monte Carlo sampler, independent draws from $\pi(X|\theta,\mathscr{D})$ or $\pi(\theta|X,\mathscr{D})$ are not essential—it is possible to substitute lines 3 and 4 of algorithm 4.1 with the corresponding Metropolis-Hasting updates. As a result, algorithm 4.2 is often used in place of algorithm 4.1. Naturally, modifications of algorithm 4.1 when only one of: parameter update or path

Algorithm 4.1 Gibbs sampler for inference for diffusion processes

1: Initialise $\theta^{(0)}$
2: **for** $n = 1, \ldots, N$ **do**
3: Draw $X^{(n)} \sim \pi(X|\theta^{(n-1)}, \mathscr{D})$ ▷ Path imputation according to $\mathbb{P}_b^{(\theta^{(n-1)})}(\cdot|\mathscr{D})$
4: Draw $\theta^{(n)} \sim \pi(\theta|X^{(n)}, \mathscr{D})$
5: **return** $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with the invariant density $\pi(\theta|\mathscr{D})$

Algorithm 4.2 Metropolis-within-Gibbs for inference for diffusion processes

1: Initialise $(\theta^{(0)}, X^{(0)})$
2: **for** $n = 1, \ldots, N$ **do**
3: Draw $X^\circ \sim q_{\theta^{(n-1)}}(X^{(n-1)}, \cdot)$ ▷ Proposal path imputation
4: Draw $U_1 \sim \texttt{Unif([0,1])}$
5: **if** $U_1 \leq \frac{\pi(\theta^{(n-1)}, X^\circ, \mathscr{D}) q_{\theta^{(n-1)}}(X^\circ, X^{(n-1)})}{\pi(\theta^{(n-1)}, X^{(n-1)}, \mathscr{D}) q_{\theta^{(n-1)}}(X^{(n-1)}, X^\circ)}$ **then**
6: $X^{(n)} \leftarrow X^\circ$
7: **else**
8: $X^{(n)} \leftarrow X^{(n-1)}$
9: Draw $\theta^\circ \sim q_{X^{(n)}}(\theta^{(n-1)}, \cdot)$ ▷ Proposal parameter update
10: Draw $U_2 \sim \texttt{Unif}([0,1])$
11: **if** $U_2 \leq \frac{\pi(\theta^\circ, X^{(n)}, \mathscr{D}) q_{X^{(n)}}(\theta^\circ, \theta^{(n-1)})}{\pi(\theta^{(n-1)}, X^{(n)}, \mathscr{D}) q_{X^{(n)}}(\theta^{(n-1)}, \theta^\circ)}$ **then**
12: $\theta^{(n)} \leftarrow \theta^\circ$
13: **else**
14: $\theta^{(n)} \leftarrow \theta^{(n-1)}$
15: **return** $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with the invariant density $\pi(\theta|\mathscr{D})$

imputation step are substituted with the corresponding Metropolis-Hastings step are also valid.

The step of path imputation—which is nothing else but a simulation of conditioned diffusion—is exactly the topic of this thesis.

4.1.1 Overview of the path imputation step

Line 3 of algorithm 4.1 prompts for independent draws from a conditioned diffusion law and (if possible) can be completed via rejection sampling on a path space, as discussed in sections 2.3 and 3.1. On the other hand, sampling from $q_\theta(X, \cdot)$ in line 3 of algorithm 4.2 amounts to drawing a proposal path in the setting of an MCMC sampler on a path space. Based on the proceedings of chapter 3, it is

clear that there are many choices available for this step. I discuss some possibilities below.

The most immediate choices of $q_\theta(X,\cdot)$ cast algorithm 4.2 in the setting of independence sampler. For exactly observed diffusions ($\mathscr{D} := \{X_{t_i}; i = 1,\ldots,K\}$) a factorisation of the target law at the times of the observations (based on the Markov property) gives $\mathbb{P}_b^{(\theta)}(\cdot|\mathscr{D}) = \bigotimes_{i=1}^K \mathbb{P}_b^{(\theta)}|_{[t_{i-1},t_i]}(\cdot|\mathscr{D}_{[t_{i-1},t_i]})$, where $\otimes$ denotes a product measure, $\mathbb{P}_b^{(\theta)}|_{[a,c]}$ denotes a restriction of the law $\mathbb{P}_b^{(\theta)}$ to the interval $[a,c]$ and $\mathscr{D}|_{[a,c]}$ denotes a restriction of the observations to those falling inside the interval $[a,c]$ (in particular $\mathscr{D}|_{[t_{i-1},t_i]} = \{X_{t_{i-1}}, X_{t_i}\}$).

For instance, if the simple diffusion bridges are chosen to be a method of drawing conditioned paths, then $q_\theta(X,\cdot)$ becomes the law $\bigotimes_{i=1}^K \mathbb{P}_b^{(\theta)}|_{[t_{i-1},t_i]}(\cdot|\mathscr{D}|_{[t_{i-1},t_i]}, \{X^{(3)} \in \mathscr{E}_X\})$ (where, with an abuse of notation, $\mathscr{E}_x$ is defined in eq. (3.5), separately for each sub-interval $[t_{i-1},t_i]$). If instead guided proposals are chosen, then $q_\theta(X,\cdot)$ becomes $\bigotimes_{i=1}^K \mathbb{P}_{b^\circ}^{(\theta)}|_{[t_{i-1},t_i]}$, defined in eq. (3.19), with $\widetilde{h}(t,x) := \mathrm{d}\mathbb{P}_b^{(\theta)}(\mathscr{D}|_{[t_{i-1},t_i]}|X_t = x)$ set on the time intervals $[t_{i-1},t_i]$, $(i = 1,\ldots,K)$. On the other hand, if noisy and partial observations are made, then guided proposals may be employed again, this time with $\widetilde{h}(t,x) := \mathrm{d}\mathbb{P}_b^{(\theta)}(\mathscr{D}|_{[t_{i-1},T]}|X_t = x)$ set on the time intervals $[t_{i-1},t_i]$, $(i = 1,\ldots,K)$.

Concrete implementations of the transition kernel $q_\theta(X^{(n-1)},\cdot)$, listed above, completely disregard the value of the previous state $X^{(n-1)}$ and therefore may be sub-optimal. In this thesis I discuss two generic techniques that aim to improve the efficiency of the path imputation step by exploiting this dependence on $X^{(n-1)}$: blocking and the preconditioned Crank-Nicolson step. I introduce the former one in chapter 5. The latter one has been introduced in section 2.7, whereas in section 4.1.6 below I show how to embed it in algorithm 4.2. Recall that in order to implement the precondition Crank-Nicolson scheme, a non-centrally parametrised version of a conditioned diffusion path sampler needs to be defined first. However, it was shown in Roberts and Stramer (2001) that the connection between non-centred parametrisation and Bayesian inference for diffusion processes goes even further then that—the former is a prerequisite for the step of parameter update—I discuss this connection in section 4.1.3.

4.1.2 Overview of the parameter update step

For the step of parameter update, notice that by Bayes theorem:

$$\pi(\theta, X, \mathscr{D}) = \pi(X|\theta, \mathscr{D})\pi(\mathscr{D}|\theta)\pi(\theta) = \frac{\mathrm{d}\mathbb{P}_b(X|\mathscr{D})}{\mathrm{d}\mathbb{Q}(X)} g_b(\mathscr{D}|\theta)\pi(\theta), \tag{4.3}$$

where $\pi(\theta)$ is a prior, $g_b(\mathscr{D}|\theta)$ is the density for observing $\mathscr{D}$ under the target measure $\mathbb{P}_b$ and $\mathbb{Q}$ denotes a proposal diffusion measure. For instance, in the case of rejection sampling on a path space $\mathbb{Q} := \mathbb{P}_\mu(\cdot|\mathscr{D})$, or in the case of guided proposals $\mathbb{Q} := \mathbb{P}_{b^\circ}$. Comparing directly with eqs. (2.10) and (3.24) a crucial observation can be made that in both of those examples the intractable density $g_b(\mathscr{D}|\theta)$ cancels with the same term appearing in the expression for $\frac{\mathrm{d}\mathbb{P}_b(X|\mathscr{D})}{\mathrm{d}\mathbb{Q}(X)}$. Therefore, employing the Metropolis-Hastings step is often a reasonable strategy.

Denote with θ_b the subset of parameters θ that enter the drift coefficient. Similarly denote with θ_σ the subset of parameters that enter the volatility coefficient. I remark that θ_b and θ_σ need not necessarily be disjoint and thus define $\theta_{b\setminus\sigma} := \theta_b \setminus \theta_\sigma$ to be the set of parameters that enter the drift, but not the volatility coefficient. It is easy to see that if $\theta_{b\setminus\sigma}$ enters the drift in a linear manner (i.e. if there exist functions $f_1(x|\theta_\sigma) : \mathbb{R}^d \to \mathbb{R}^d$ and $f_2(x|\theta_\sigma) : \mathbb{R}^d \to \mathbb{R}^{d\times d_{b\setminus\sigma}}$, with $d_{b\setminus\sigma}$ denoting the dimension of $\theta_{b\setminus\sigma}$, such that $b(x) = f_1(x|\theta_\sigma) + f_2(x|\theta_\sigma)\theta_{b\setminus\sigma}$) then, using arguments analogous to van der Meulen and Schauer (2017a, §3.3) it can be shown that the likelihood $\pi(\theta_{b\setminus\sigma}|\theta_\sigma, X^{(n)}, \mathscr{D})$ is conjugate to Gaussian priors and thus any subset of the parameters $\theta_{b\setminus\sigma}$ can be sampled directly from the full conditional distribution. I illustrate this on the example of the Ornstein-Uhlenbeck process in example 4.1.1 below. I also show in section 7.4.2.2 how to extend this idea to a particular type of hypoelliptic diffusions. Otherwise, if the relationship of $\theta_{b\setminus\sigma}$ is not linear, because the form of $\pi(\theta_{b\setminus\sigma}|\theta_\sigma, X^{(n)}, \mathscr{D})$ is known, implementing the Metropolis-Hastings algorithm with any reasonable kernel q (say a random walk) is simple. Surprisingly, the step of updating $\pi(\theta_\sigma|\theta_{b\setminus\sigma}, X^{(n)}, \mathscr{D})$ is ill-defined and thus modifications to algorithms 4.1 and 4.2 are essential to correct this fault.

4.1.3 Mutual singularity of measures

Roberts and Stramer (2001) showed emphatically the degeneracy of the parameter update step $\pi(\theta_\sigma|\theta_{b\setminus\sigma}, X^{(i)}, \mathscr{D})$. Notice that conditionally on the path $X^{(n)}$, the

diffusion coefficient Γ_θ is uniquely identified via the relation:

$$[X,X]_t = \int_0^t \Gamma_\theta(X_s)\,\mathrm{d}s,$$

where $[\cdot,\cdot]$ denotes the quadratic variation and therefore

$$\pi(\theta_\sigma|\theta_{b\setminus\sigma}, X^{(n)}, \mathscr{D}) \propto \mathbb{1}_{\{\vartheta\in\mathbb{R}^{d_\sigma}:\sigma_\vartheta\sigma_\vartheta^T=\Gamma\}}(\theta_\sigma),$$

where d_σ is the dimension of the θ_σ vector. Consequently, once θ_σ is set, algorithms 4.1 and 4.2 are never allowed to update it (and if the integrals in eq. (2.3) are approximated with Riemann sums, then Roberts and Stramer (2001) showed that the mixing of the chain updating parameter θ_σ slows down quadratically with the number of imputed points). In other words, the diffusion laws $\mathbb{P}_b$ with different volatility coefficients are mutually singular.

4.1.4 Non-centred parametrisation

As shown in Roberts and Stramer (2001), non-centred parametrisation of the diffusion path sampler allows to remove the aforementioned degeneracy and is thus indispensable in the setting of Bayesian inference for diffusion processes. As in section 2.6, I use $(\Omega^*, \mathscr{F}^*, \mathbb{Q}^*)$ to denote a non-centred probability space and let $\Psi_\theta : \Omega^* \to \Omega$ be a deterministic function for which the pushforward measure $(\mathbb{Q}^*)_\#(\Psi_\theta)$ coincides with the proposal measure $\mathbb{P}_0^{(\theta)}$, $\theta \in \Theta$. The goal is to define a Markov chain $\{(\theta^{(n)}, X^{(n)}, W^{(n)}); n = 0,\ldots,2N\}$ whose invariant distribution is characterised by the following two properties:

- The invariant density of the marginal chain $\{(\theta^{(n)}, X^{(n)}); n = 0,\ldots,2N\}$ is equal to $\pi(\theta, X|\mathscr{D})$.
- For each $n = 0,\ldots,2N$, the following identity holds: $\Psi_{\theta^{(n)}}(W^{(n)}) = X^{(n)}$.

Such parametrisation can be defined readily by making minimal changes to algorithm 4.2, so that instead of directly updating path X, it is only every calculated as a by-product from the pair of (W,θ)—the fact that $(\mathbb{Q}^*)_\#(\Psi_\theta)$ coincides with $\mathbb{P}_0^{(\theta)}$ guarantees that the algorithm is still valid. An updated procedure is summarised in algorithm 4.3.

Algorithm 4.3 Non-centrally parametrised inference for diffusion processes

Initialise $(\theta^{(0)}, W^{(0)})$, set $X^{(0)} \leftarrow \Psi_{\theta^{(0)}}(W^{(0)})$
for $n = 0, \ldots, N-1$ **do**
  Draw $W^\circ \sim q_{\theta^{(2n)}}(W^{(2n)}, \cdot)$ ▷ Proposal path imputation
  Set $X^\circ \leftarrow \Psi_{\theta^{(2n)}}(W^\circ)$
  Draw $U_1 \sim \texttt{Unif([0,1])}$
  if $U_1 \leq \frac{\pi(\theta^{(2n)}, X^\circ, \mathscr{D}) q_{\theta^{(2n)}}(W^{(2n)}, W^\circ)}{\pi(\theta^{(2n)}, X^{(2n)}, \mathscr{D}) q_{\theta^{(2n)}}(W^{(2n)}, W^\circ)}$ **then**
    $(\theta^{(2n+1)}, X^{(2n+1)}, W^{(2n+1)}) \leftarrow (\theta^{(2n)}, X^\circ, W^\circ)$
  else
    $(\theta^{(2n+1)}, X^{(2n+1)}, W^{(2n+1)}) \leftarrow (\theta^{(2n)}, X^{(2n)}, W^{(2n)})$
  Draw $\theta^\circ \sim q_{W^{(2n+1)}}(\theta^{(2n+1)}, \cdot)$ ▷ Proposal parameter update
  Set $X^\circ \leftarrow \Psi_{\theta^\circ}(W^{(2n+1)})$
  Draw $U_2 \sim \texttt{Unif}([0,1])$
  if $U_2 \leq \frac{\pi(\theta^\circ, X^\circ, \mathscr{D}) q_{W^{(2n+1)}}(\theta^\circ, \theta^{(2n+1)})}{\pi(\theta^{(2n+1)}, X^{(2n+1)}, \mathscr{D}) q_{W^{(2n+1)}}(\theta^{(2n+1)}, \theta^\circ)}$ **then**
    $(\theta^{(2n+2)}, X^{(2n+2)}, W^{(2n+2)}) \leftarrow (\theta^\circ, X^\circ, W^{(2n+1)})$
  else
    $(\theta^{(2n+2)}, X^{(2n+2)}, W^{(2n+2)}) \leftarrow (\theta^{(2n+1)}, X^{(2n+1)}, W^{(2n+1)})$
return $\{\theta^{(2n)}; n = 0, \ldots, N\}$ ▷ Markov chain with the invariant density $\pi(\theta|\mathscr{D})$

In particular, notice that unlike $\pi(\theta^\circ, X^{(n)}, \mathscr{D})$ in line 11 of algorithm 4.2, which is non-zero only for $\theta^\circ_\sigma \in \{\vartheta \in \mathbb{R}^{d_\sigma} : \sigma_\vartheta \sigma_\vartheta^T = \Gamma_{\theta^{(n)}}\}$ and renders algorithm 4.2 degenerate, $\pi(\theta^\circ, X^\circ, \mathscr{D})$ in line 13 of algorithm 4.3 is well defined for all $\theta^\circ \in \Theta$, because X° is updated alongside θ°.

4.1.5 Examples

Example 4.1.1. Suppose that $\mathscr{D} := \{X_{t_i}; i = 1, \ldots, K\}$ and that the underlying process is modelled with the Ornstein-Uhlenbeck process:

$$\mathrm{d}X_t = [\theta^{[1]} - \theta^{[2]} X_t]\,\mathrm{d}t + \theta^{[3]}\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T]. \tag{4.4}$$

The Ornstein-Uhlenbeck process is one of those rare diffusions for which transition densities are available in closed forms. Nonetheless, for didactic purposes I will not resort to them and instead show how to perform inference via data augmentation. For the step of path imputation I will employ rejection sampling on a path space and whenever appropriate approximate the integrals with left-Riemann sums.

First, Lamperti transformation takes the following, simple form:

$$\eta_\theta(x) := \frac{x}{\theta^{[3]}},$$

and thus the Lamperti transformed diffusion $Y := \{\eta_\theta(X_t);\, t \in [0,T]\}$ solves:

$$\mathrm{d}Y_t = \left[\frac{\theta^{[1]}}{\theta^{[3]}} - \theta^{[2]} Y_t\right]\mathrm{d}t + \mathrm{d}W_t, \quad Y_0 = y_0 := \frac{x_0}{\theta^{[3]}}, \quad t \in [0,T]. \tag{4.5}$$

The non-centred probability space is defined similarly to the one in example 2.6.1. Let the non-centred process $Z := \{Z^{[i]};\, i = 0,\ldots,K-1\}$ be defined on a probability space $(\Omega^*, \mathscr{F}^*, \mathbb{Q}^*)$ and take values in a measurable space $(\mathscr{X}^*, \Sigma)$, where $\Omega^* \equiv \mathscr{X}^*$ are defined as:

$$\Omega^* \equiv \mathscr{X}^* := \prod_{i=0}^{K-1} \mathcal{C}([0,\Delta_i];\mathbb{R}), \text{ where } \Delta_i := t_{i+1} - t_i, \quad t_0 := 0,$$

where $\prod$ denotes a Cartesian product, $\Sigma \equiv \mathscr{F}^* = \mathscr{B}(\mathscr{X}^*)$ are Borel-σ-algebras and $\mathbb{Q}^*$ is the product measure of K independent 0–0 Brownian bridges on $[0,\Delta_i]$, $(i = 0,\ldots,K-1)$. Function

$$\Psi_\theta : \prod_{i=0}^{K-1} \mathcal{C}([0,\Delta_i];\mathbb{R}) \to \mathcal{C}([0,T];\mathbb{R}),$$

takes the form analogous to (2.20):

$$\begin{aligned} \Psi_\theta(Z) &:= \left\{\sum_{i=0}^{K-1} \Psi_\theta^{[i]}(Z)\mathbb{1}_{(t_i,t_{i+1}]}(t);\, t \in [0,T]\right\}, \quad \text{where} \\ \Psi_\theta^{[i]}(Z) &:= \left\{Z^{[i]}_{t-t_i} + \frac{x_i}{\theta^{[3]}}(\theta)\left(1 - \frac{t-t_i}{\Delta_i}\right) + \frac{x_{i+1}}{\theta^{[3]}}\frac{t-t_i}{\Delta_i};\, t \in [t_i, t_{i+1}]\right\}, \end{aligned} \tag{4.6}$$

and $x_i := X_{t_i}$, $(i = 0,\ldots,K)$. To put it simply, the non-centred process Z consists of K independent 0–0 Brownian bridges, which Ψ_θ defined in eq. (4.6) linearly translates in such a way that the end-points at which the bridges are anchored agree with the Lamperti transformed observation set $\mathscr{D}$. This is illustrated graphically in fig. 4.1.

By the Markov property, a path imputation step can be performed independently on each sub-interval $[t_i, t_{i+1}]$, $(i = 0,\ldots,K-1)$. Correspondingly, the likelihood function for the path Y factorises at the observation times (it follows readily after combining eqs. (2.8) and (2.10)) and is given by:

$$\pi(Y|\theta,\mathscr{D}) \propto \prod_{i=0}^{K-1} \exp\left\{-\frac{(\theta^{[2]})^2}{2}\int_{t_i}^{t_{i+1}} \left(Y_t - \frac{\theta^{[1]}}{\theta^{[2]}\theta^{[3]}}\right)^2 \mathrm{d}t\right\} \leq 1.$$

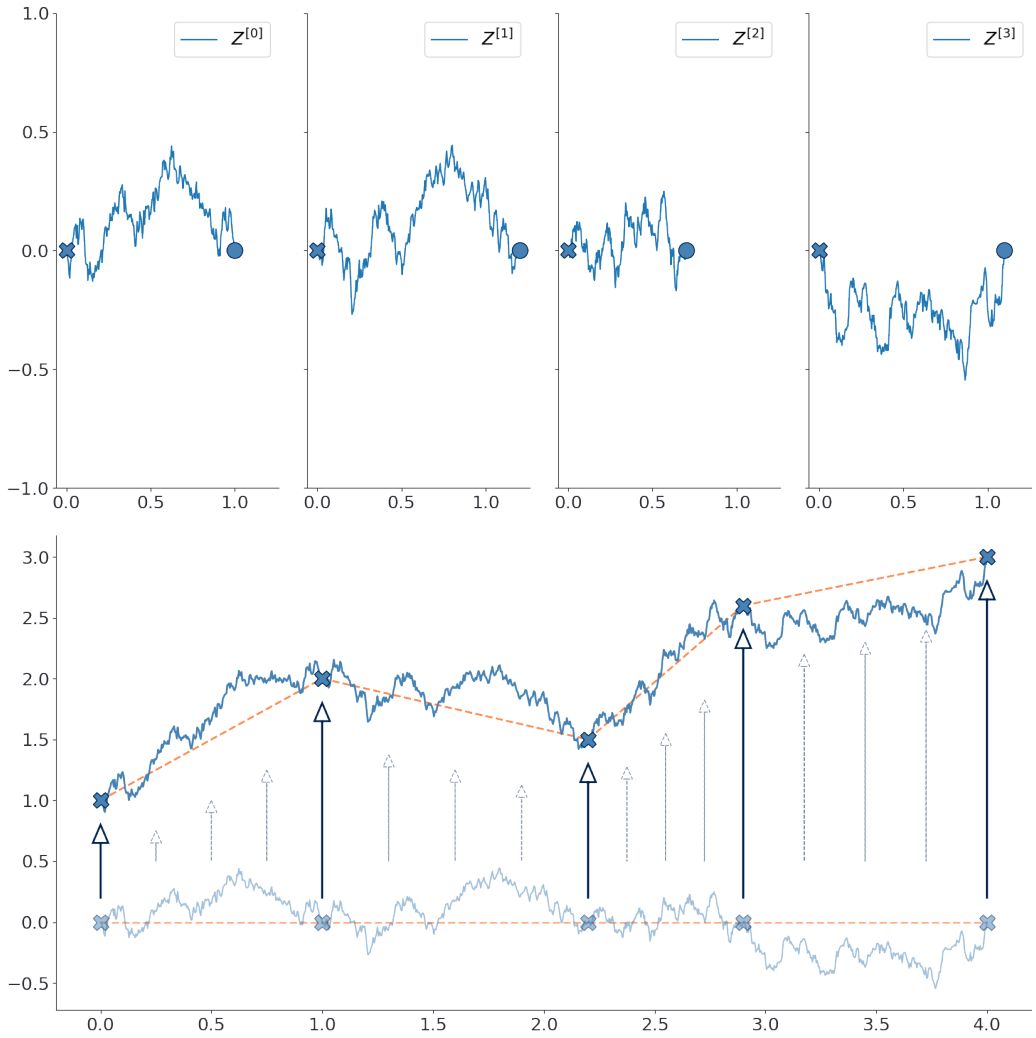


Figure 4.1: Illustration of function Ψ_θ linearly translating non-centrally parametrised bridges to proposal bridges. Top row gives samples of the non-centred process on 4 intervals. These samples are then linearly translated in the bottom plot so that their end-points agree with the Lamperti transformed observations.

The step of path imputation can therefore be completed with the following algorithm 4.4, which is a concrete, non-centrally parametrised implementation of line 3 of algorithm 4.1

To update parameter θ, notice that by eqs. (2.3), (2.10) and (4.3) the joint like-

Algorithm 4.4 Non-centred path imputation at the n^{th} iteration of a Markov chain

for $i = 0, \ldots, K-1$ **do**
  Set `accepted` $\leftarrow$ `False`
  repeat
    Draw $Z^{\circ[i]} \sim \mathbb{W}^*[0, \Delta_i]$ on ▷ 0–0 Brownian bridge on $[0, \Delta_i]$
    Set $Y^\circ|_{[t_i, t_{i+1}]} \leftarrow \Psi_\theta^{[i]}(Z^{\circ[i]})$
    Draw $E \sim$ `Exp(1)`
    if $E \geq \frac{(\theta^{[2]})^2}{2} \int_{t_i}^{t_{i+1}} \left(Y_t^\circ - \frac{\theta^{[1]}}{\theta^{[2]}\theta^{[3]}} \right)^2 - \left(Y_t^{(n)} - \frac{\theta^{[1]}}{\theta^{[2]}\theta^{[3]}} \right)^2 \mathrm{d}t$ **then**
      `accepted` $\leftarrow$ `True`
  until `accepted`
Set $(\theta^{(n+1)}, Y^{(n+1)}, Z^{(n+1)}) \leftarrow (\theta^{(n)}, Y^\circ, Z^\circ)$
return $(\theta^{(n+1)}, Y^{(n+1)}, Z^{(n+1)})$ ▷ New state of a Markov chain

lihood can be written as:

$$\begin{aligned}\pi(\theta, X, \mathscr{D}) &= \pi(\theta) g_b(\mathscr{D}) \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}}{\mathrm{d}\mathbb{P}_\mu^{(\theta)}}(X|\mathscr{D}) = \pi(\theta) \widetilde{g}_0(\mathscr{D}) \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(X) \\ &= \pi(\theta)(2\pi)^{-\frac{K}{2}} (\theta^{[3]})^{-K} \exp\left\{ -\frac{1}{2(\theta^{[3]})^2} \left[(\theta^{[1:2]} - \Lambda^{-1}\mathscr{W})^T \Lambda (\theta^{[1:2]} - \Lambda^{-1}\mathscr{W}) \right.\right. \\ &\qquad \left.\left. - \mathscr{W}^T \Lambda^{-1} \mathscr{W} + \sum_{i=0}^{K-1} (x_{i+1} - x_i)^2 \right] \right\}, \end{aligned} \tag{4.7}$$

where

$$\mathscr{W} := \int_0^T \begin{pmatrix} 1 \\ -X_s \end{pmatrix} \mathrm{d}X_s, \qquad \Lambda := \int_0^T \begin{pmatrix} 1 & -X_s \\ -X_s & X_s^2 \end{pmatrix} \mathrm{d}s,$$

and $g_b(\mathscr{D})$ and $\widetilde{g}_0(\mathscr{D})$ denote the density functions for observing $\mathscr{D}$ under the target law $\mathbb{P}_b^{(\theta)}$ (induced by eq. (4.4)) and the proposal law $\mathbb{P}_0^{(\theta)}$ (induced by a scaled Brownian motion $\theta^{[3]}W$) respectively. If the prior factorises as $\pi(\theta) := \pi(\theta^{[1:2]})\pi(\theta^{[3]})$, where $\pi(\theta^{[1:2]})$ is a Gaussian prior with mean μ_{pr} and precision Λ_{pr}, then one full-conditional distribution follows easily:

$$\pi(\theta^{[1:2]}|\theta^{[3]}, X, \mathscr{D}) \sim \mathtt{Gsn}\left(\left[\frac{1}{(\theta^{[3]})^2}\Lambda + \Lambda_{pr} \right]^{-1} \left[\frac{1}{(\theta^{[3]})^2}\mathscr{W} + \Lambda_{pr}\mu_{pr} \right], \left[\frac{1}{(\theta^{[3]})^2}\Lambda + \Lambda_{pr} \right]^{-1} \right),$$

and thus update of $\theta^{[1:2]}$ is clear. For updating $\theta^{[3]}$ the Metropolis-Hastings algorithm is necessary, and although the form of the joint density given in eq. (4.7) can be used to complete this step, it is preferable to derive an expression which does

not involve any stochastic integrals. They enter eq. (4.7) through $\mathscr{W}$, and due to their non-zero quadratic variation have higher variance than their Lebesgue counterparts.

From eqs. (2.8), (2.10) and (4.3) it follows that the joint density can be decomposed as follows:

$$\begin{aligned}\pi(\theta, X, \mathscr{D}) &= \pi(\theta) g_b(\mathscr{D}) \frac{g_0(\mathscr{D}_\theta)}{g_\alpha(\mathscr{D}_\theta)} \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(Y) = \pi(\theta) |\det(\sigma_\theta)|^{-K} g_0(\mathscr{D}_\theta) \frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(Y) \\ &= \pi(\theta)(2\pi)^{-\frac{K}{2}} (\theta^{[3]})^{-K} \exp\left\{ A_\theta(Y_T) - A_\theta(Y_0) - \int_0^T \varphi_\theta(Y_s)\mathrm{d}s \right\},\end{aligned} \tag{4.8}$$

where

$$A_\theta(y) := \frac{\theta^{[1]}}{\theta^{[3]}} y - \frac{\theta^{[2]}}{2} y^2, \quad \text{and} \quad \varphi_\theta(y) := \frac{1}{2}\left[\left(\frac{\theta^{[1]}}{\theta^{[3]}} - \theta^{[2]} y \right)^2 - \theta^{[2]} \right],$$

where $\mathscr{D}_\theta$ denotes the Lamperti transformed data, and where $g_0(\mathscr{D}_\theta)$ and $g_\alpha(\mathscr{D}_\theta)$ denote the density functions for observing $\mathscr{D}_\theta$ under the Lamperti transformed target law $\mathbb{P}_\alpha^{(\theta)}$ (induced by eq. (4.5)) and under the proposal law $\mathbb{P}_0^{(\theta)}$ (of a Brownian motion) respectively. I additionally used $g_b(\mathscr{D}) = |\det(\sigma_\theta)|^{-K} g_\alpha(\mathscr{D}_\theta)$, which follows by the change of variables formula. A parameter update step can now be performed; its summary is given in algorithm 4.5. Mind the notation $^{(n)[i]}$—a number in the parenthesis indicates an iteration of a Markov chain, whereas the number in the square bracket indicates an index of a coordinate.

Example 4.1.2. Consider again the setting from example 4.1.1 of the exactly and discretely observed Ornstein-Uhlenbeck process. Let me illustrate how to repeat the inference via data-augmentation using guided proposals instead of rejection sampling on a path space. For simplicity of exposition I set the auxiliary process $\widetilde{X}$ to be given by a scaled Brownian motion $\theta^{[3]} W$. This results in the following proposals:

$$\mathrm{d}X_t = \left[\theta^{[1]} - \theta^{[2]} X_t + \frac{x_{i+1} - X_t}{t_{i+1} - t} \right] \mathrm{d}t + \theta^{[3]} \mathrm{d}W_t, \quad X_{t_i} = x_i, \quad t \in [t_i, t_{i+1}], \tag{4.9}$$

Algorithm 4.5 Parameter update at the n^{th} step of a Markov chain

1: Set $\mathscr{W} \leftarrow \int_0^T \begin{pmatrix} 1 \\ -X_s^{(n)} \end{pmatrix} \mathrm{d}X_s^{(n)}$
2: Set $\Lambda \leftarrow \int_0^T \begin{pmatrix} 1 & -X_s^{(n)} \\ -X_s^{(n)} & (X_s^{(n)})^2 \end{pmatrix} \mathrm{d}s$
3: Draw $\theta^{\circ[1:2]} \sim \mathtt{Gsn}\left(\left[\frac{1}{(\theta^{(n)[3]})^2}\Lambda + \Lambda_{pr}\right]^{-1}\left[\frac{1}{(\theta^{(n)[3]})^2}\mathscr{W} + \Lambda_{pr}\mu_{pr}\right], \left[\frac{1}{(\theta^{(n)[3]})^2}\Lambda + \Lambda_{pr}\right]^{-1}\right)$
4: Draw $\theta^{\circ[3]} \sim q(\theta^{(n)[3]}, \cdot)$ ▷ Say, a random walk
5: Set $Y^\circ \leftarrow \Psi_{\theta^\circ}(Z^{(n)})$
6: Set $\theta^{*[1:2]} \leftarrow \theta^{\circ[1:2]}$, and $\theta^{*[3]} \leftarrow \theta^{(n)[3]}$ ▷ Update of $\theta^{\circ[1:2]}$ is always accepted
7: Set $Y^* \leftarrow \Psi_{\theta^*}(Z^{(n)})$
8: Set $\mathtt{ll}^* \leftarrow \log(\pi(\theta^{*[3]})) - K\log(\theta^{*[3]}) + A_{\theta^*}(Y_T^*) - A_{\theta^*}(Y_0^*) - \int_0^T \varphi_{\theta^*}(Y_s^*)\mathrm{d}s$
9: Set $\mathtt{ll}^\circ \leftarrow \log(\pi(\theta^{\circ[3]})) - K\log(\theta^{\circ[3]}) + A_{\theta^\circ}(Y_T^\circ) - A_{\theta^\circ}(Y_0^\circ) - \int_0^T \varphi_{\theta^\circ}(Y_s^\circ)\mathrm{d}s$
10: Draw $E \sim \mathtt{Exp(1)}$
11: **if** $E \geq -\mathtt{ll}^\circ + \mathtt{ll}^*$ **then**
12: Set $(\theta^{(n+1)}, Y^{(n+1)}, Z^{(n+1)}) \leftarrow (\theta^\circ, Y^\circ, Z^{(n)})$
13: **else**
14: Set $(\theta^{(n+1)}, Y^{(n+1)}, Z^{(n+1)}) \leftarrow (\theta^*, Y^*, Z^{(n)})$ ▷ Reject update of $\theta^{[3]}$, but accept $\theta^{[1:2]}$
15: **return** $(\theta^{(n+1)}, Y^{(n+1)}, Z^{(n+1)})$ ▷ New state of a Markov chain

$(i = 0, \ldots, K-1)$. The likelihood function for an imputed path X follows from eq. (3.24) and is equal to:

$$\pi(X|\theta, \mathscr{D}) \propto \prod_{i=0}^{K-1} \exp\left\{\int_{t_i}^{t_{i+1}} (\theta^{[1]} - \theta^{[2]}X_s)\frac{x_{i+1} - X_s}{t_{i+1} - s}\,\mathrm{d}s\right\}.$$

A suitable non-centred probability space can be defined as in example 4.1.1, with the exception that the law $\mathbb{Q}^*$ is now induced by the product measure of K independent laws induced by standard Brownian motion on $[0, \Delta_i]$, $(i = 0, \ldots, K-1)$. Function

$$\Psi_\theta : \prod_{i=0}^{K-1} \mathcal{C}([0, \Delta_i]; \mathbb{R}) \to \mathcal{C}([0, T]; \mathbb{R}),$$

(with $\prod$ denoting the Cartesian product) has a deceptively similar form as before:

$$\Psi_\theta(Z) := \left\{\sum_{i=0}^{K-1} \Psi_\theta^{[i]}(Z)\mathbb{1}_{(t_i, t_{i+1}]}(t);\ t \in [0, T]\right\}, \tag{4.10}$$

but now, $\Psi_\theta^{[i]}(Z)$ are defined as solutions to eq. (4.9) when the driving Brownian motion W takes a path-value $Z^{[i]}$, $(i = 0, \ldots, K-1)$ (hence the need for assumption A2, mentioned in chapter 2). For instance, when using the Euler-Maruyama

scheme, the increments of $Z^{[i]}$: $\{Z^{[i]}_{s_{j+1}} - Z^{[i]}_{s_j}; j = 0,\dots,m\}$ are used to recursively compute $X|_{[t_i,t_{i+1}]}$ via:

$$X_{s_{j+1}} = \left[\theta^{[1]} - \theta^{[2]} X_{s_j} + \frac{x_{i+1} - X_{s_j}}{t_{i+1} - s_j}\right](s_{j+1} - s_j) + \theta^{[3]}(Z^{[i]}_{s_{j+1}} - Z^{[i]}_{s_j}), \quad j = 0,\dots,m,$$

where $t_i = s_0 < \cdots < s_{m+1} = t_{i+1}$, $(i = 0,\dots,K-1)$. $\Psi^{[i]}_\theta(Z)$ is therefore given exactly by the recursive mapping above, taking the driving noise Z and mapping it to a proposal path X. Algorithm 4.6 summarises the path imputation step, with $G^{[i]}_\theta$ defined as

$$G^{[i]}_\theta(s,x) := (\theta^{[1]} - \theta^{[2]} x)\frac{x_{i+1} - x}{t_{i+1} - s}.$$

Algorithm 4.6 Path imputation for guided proposals at the n^{th} step of a Markov chain

for $i = 0,\dots,K-1$ **do**
  Draw $Z^{\circ[i]} \sim \mathbb{W}^*$ ▷ Brownian motion on $[0, \Delta_i]$
  Set $X^\circ|_{[t_i,t_{i+1}]} \leftarrow \Psi^{[i]}_{\theta^{(i)}}(Z^\circ)$ ▷ the Euler-Maruyama scheme
  Draw $E \sim \texttt{Exp(1)}$
  if $E \geq -\int_{t_i}^{t_{i+1}} G^{[i]}_{\theta^{(n)}}(s, X^\circ_s)\,\mathrm{d}s + \int_{t_i}^{t_{i+1}} G^{[i]}_{\theta^{(n)}}(s, X^{(n)}_s)\,\mathrm{d}s$ **then**
    Set $(X^{(n+1)}|_{[t_i,t_{i+1}]}, Z^{(n+1)[i]}) \leftarrow (X^\circ|_{[t_i,t_{i+1}]}, Z^{\circ[i]})$
  else
    Set $(X^{(n+1)}|_{[t_i,t_{i+1}]}, Z^{(n+1)[i]}) \leftarrow (X^{(n)}|_{[t_i,t_{i+1}]}, Z^{(n)[i]})$
Set $\theta^{(n+1)} \leftarrow \theta^{(n)}$
return $(\theta^{(n+1)}, Y^{(n+1)}, Z^{(n+1)})$ ▷ New state of a Markov chain

Parameter update can be performed in a similar manner as in example 4.1.1. In particular, the conditional law $\pi(\theta^{[1:2]}|\theta^{[3]}, X, \mathscr{D})$ remains unchanged and even (4.8) could in principle be reused. However, now it is also possible to use a different representation for the joint density:

$$\begin{aligned}\pi(\theta, X, \mathscr{D}) &= \pi(\theta) g_b(\mathscr{D}) \frac{\mathrm{d}\mathbb{P}_b(X|\mathscr{D})}{\mathrm{d}\mathbb{P}_{b^\circ}(X)} \\ &= \pi(\theta)(2\pi)^{-\frac{K}{2}}(\theta^{[3]})^{-K} \prod_{i=0}^{K-1} \exp\left\{-\frac{\sum_{i=0}^{K-1}(x_{i+1} - x_i)^2}{2(\theta^{[3]})^2}\right\} \exp\left\{\int_{t_i}^{t_{i+1}} G^{[i]}_\theta(X_s)\,\mathrm{d}s\right\}.\end{aligned}$$

Modifications to algorithm 4.5 for accommodating guided proposals follow trivially.

The procedures developed in examples 4.1.1 and 4.1.2, even though aimed at a simple problem of inference for the Ornstein-Uhlenbeck process, differ very little from the corresponding steps that would have needed to be taken under a more elaborate diffusion model or in another observational setting (which does not mean that the changes that would have needed to be made are trivial). In general, the step of updating $\theta_{b\setminus\sigma}$, which in examples 4.1.1 and 4.1.2 was performed via direct sampling from $\pi(\theta_{b\setminus\sigma}|\theta_\sigma, X, \mathscr{D})$, might need to be substituted with the Metropolis-Hastings step, as per algorithm 4.3. It can thus be performed in an analogous way to how the update of θ_σ was completed in examples 4.1.1 and 4.1.2. Computations of G_θ, φ_θ, α_θ etc. will also change accordingly.

Under different observational schemes or, for guided proposals, more elaborate choices of $\widetilde{X}$, $g_0(\mathscr{D})$ or $\widetilde{g}_0(\mathscr{D})$ might also change their forms. In fact, for guided proposals, if $\widetilde{X}$ is complicated enough, then computations of $\widetilde{\tau}$ and $\widetilde{H}$ from their closed form expressions might be inefficient and it might be preferable to use schemes from chapter 6 to find their accurate approximations instead. In chapter 7, where different observational scheme is considered, I also alter the proposal process accordingly. Additionally, the algorithms admit certain extensions, such as the preconditioned Crank-Nicolson scheme (see section 4.1.6) or blocking (see chapter 5), that slightly alter the form of the protocol. Beyond those remarks, the methodology of examples 4.1.1 and 4.1.2 is representative of the general proceedings for Bayesian inference for diffusion processes via data augmentation.

4.1.6 Preconditioned Crank-Nicolson scheme

The preconditioned Crank-Nicolson scheme, discussed in section 2.7, has been introduced in the context of MCMC sampler on a path space as a way of making local moves on that space. Algorithm 4.3 is formulated in such a way that it is straightforward to see how the preconditioned Crank-Nicolson scheme can be embedded inside the step of path imputation. Notice that line 3 of algorithm 4.3 is all that needs to change. In particular, $W^\circ \sim q_{\theta^{(2n)}}(W^{(2n)}, \cdot)$ takes the form given in algorithm 4.7. Whether $\mathbb{W}^*$ denotes a law of a 0–0 Brownian bridge, or standard Brownian motion depends on which path sampling algorithm is being employed.

Algorithm 4.7 Preconditioned Crank-Nicolson update at the $2n^{th}$ step of a Markov chain

1: Draw $W^\circ \sim \mathbb{W}^*$ ▷ 0–0 Brownian bridge, or standard Brownian motion
2: Set $W^\circ \leftarrow \sqrt{\lambda}W^\circ + \sqrt{1-\lambda}W^{(2n)}$
3: **return** W° ▷ Locally perturbed driving noise

4.2 Exact Bayesian inference for diffusion processes

In section 3.1 I described the *exact rejection sampling on a path space*, due to Beskos and Roberts (2005); Beskos et al. (2006, 2008), which is a method of sampling diffusion bridges that does not introduce any discretisation errors. In the setting of Bayesian inference for diffusion processes, formulation of algorithm 4.3 hints that imbuing a path imputation step with the *exactness* property might be a wasteful exercise, as the property is— seemingly—immediately lost in a subsequent step of updating parameters (the joint likelihood function $\pi(\theta, X, \mathscr{D})$ involves expressions with integrals over path functionals and approximating those with left-Riemann sums introduces discretisation error). Surprisingly, as shown in Beskos et al. (2006) and Sermaidis et al. (2013), a careful formulation of the Bayesian inference algorithm need not be burdened with errors due to approximations. Beskos et al. (2006) discuss a number of interesting possibilities for exact inference algorithms for diffusion processes; however, it is the methodology of Sermaidis et al. (2013) that is the most relevant to the developments of this thesis.

Conceptually, Sermaidis et al. (2013) work with the same Markov chain $\{(\theta^{(n)}, X^{(n)}); n = 0, \ldots, N\}$ whose invariant density is given by $\pi(\theta, X|\mathscr{D})$, although a concrete implementation uses a slightly different chain. Recall from algorithms 3.1 and 3.2 that the exact rejection sampler on a path space prompts for simulation of a unit intensity Poisson point process Φ on $[0, T] \times [0, l^*(\Upsilon)]$ and an acceptance decision is made solely on the basis of revealing a path of the Lamperti transformed diffusion Y at the times $\{\chi_j ; j = 1, \ldots, \varkappa\}$. It turns out that if this algorithm is used for the path imputation step, post-acceptance there is no need for revealing Y at any additional time points and instead, working directly with $\{Y_{\chi_j} ; j = 1, \ldots, \varkappa\}$ is possible.

Under the exact observation scheme $\mathscr{D} := \{X_{t_i}; i = 1,\ldots,K\}$, Sermaidis et al. (2013) define a Markov chain $\{\theta^{(n)}, \mathscr{S}^{(n)}; n = 0,\ldots,N\}$, where

$$\mathscr{S} := \bigcup_{i=0}^{K-1} \left\{ \left\{ \mathrm{Z}^{[i]}_{\chi_j^{(i)}}; j = 1,\ldots,\chi_i \right\} \cup \{l^*(\Upsilon_i)\} \right\},$$

(with $\mathrm{Z} := (\mathrm{Z}^{[i]})_{i=0}^{K-1}$ and $\mathrm{Z}^{[i]}$ denoting the non-centred process on an interval $[t_i, t_{i+1}]$, $i = 0,\ldots,K-1$) is a finite dimensional summary information about the imputed path Y, obtained from collating all random variables produced by the exact rejection sampler on a path space in the process of generating a single, accepted sample of Y. I refer to $\mathscr{S}$ as a surrogate path or a surrogate variable. $\{\theta^{(n)}, \mathscr{S}^{(n)}; n = 0,\ldots,N\}$ is constructed by modifying a non-centrally parametrised instance of the Metropolis-within-Gibbs algorithm 4.3 to accommodate a change from X to $\mathscr{S}$. A path imputation step now boils down to drawing a surrogate path $\mathscr{S}$ via the non-centrally parametrised exact rejection sampler on a path space (see section 3.1). To update parameter θ, the joint density for the parameter θ, the surrogate variable $\mathscr{S}$ and the observations $\mathscr{D}$ is derived in Sermaidis et al. (2013, Theorem 1):

$$\begin{aligned}\pi(\theta, \mathscr{S}, \mathscr{D}) = \pi(\theta) \exp\{A_\theta(\Psi_\theta(\mathrm{Z})_T) - A_\theta(\Psi_\theta(\mathrm{Z})_0) - (l_{*\theta} - 1)T\} \\ \cdot \prod_{i=1}^{K} \left\{ D_\theta(x_i) \exp\{-l^*_\theta(\Upsilon_i)\Delta_i\} \mathscr{N}_{\Delta_i}[\eta_\theta^{-1}(x_{i+1}) - \eta_\theta^{-1}(x_i)] \cdot [l^*_\theta(\Upsilon_i)]^{\chi_i} \right. \\ \left. \cdot \prod_{j=1}^{\chi_i} \left[1 - \phi_\theta(\Psi_\theta^{[i]}(\mathrm{Z})_{\chi_j^{(i)}}) / l^*_\theta(\Upsilon_i) \right] \right\},\end{aligned} \tag{4.11}$$

where $\Delta_i := t_{i+1} - t_i$ and $\mathscr{N}_\Sigma(x)$ denotes a Gaussian pdf with mean 0 and variance Σ, evaluated at x and where $D_\theta(x) := |\det(\sigma_\theta(x))|^{-1}$. See the proof of Sermaidis et al. (2013, Theorem 1) for derivation details. The Metropolis-Hastings algorithm can thus be employed to update parameter θ, leading to a single sweep of the entire Metropolis-within-Gibbs algorithm consisting of a draw from $\mathscr{S} \sim \pi(\mathscr{S}|\theta^{(n-1)}, \mathscr{D})$, followed by a proposal θ° sampled from some transition kernel q and a decision to accept it, taken with the probability:

$$a(\theta^{(n-1)}, \theta^\circ) := 1 \wedge \frac{\pi(\theta^\circ, \mathscr{S}, \mathscr{D}) q(\theta^\circ, \theta^{(n-1)})}{\pi(\theta^{(n-1)}, \mathscr{S}, \mathscr{D}) q(\theta^{(n-1)}, \theta^\circ)}. \tag{4.12}$$

A complete procedure is summarised in algorithm 4.8 below.

Algorithm 4.8 Exact inference for diffusion processes via data augmentation

Initialise $\theta^{(0)}$
for $n = 1, \ldots, N$ **do**
  for $i = 0, \ldots, K-1$ **do**
    repeat ▷ Path imputation on $[t_i, t_{i+1}]$ via exact rejection sampling
      Set `accepted` ← `True`
      Draw $\Upsilon^{[i]}$ ▷ Layer information, see section 3.1
      Draw $\Phi := \{(\chi_k, \psi_k); k = 1, \ldots, \varkappa\} \sim \mathbb{L}$ on $[t_i, t_{i+1}] \times [0, l^*_{\theta^{(n-1)}}(\Upsilon_i)]$
      Draw $W^\circ \sim \mathbb{P}_0(\cdot|\mathscr{D}|_{[t_i, t_{i+1}]} \cap \Upsilon_i)$ at times $\{\chi_k ; k = 1, \ldots, \varkappa\}$
      Set $Y^\circ_{\chi_k} \leftarrow [\Psi^{[i]}_{\theta^{(n-1)}}(W^\circ)]_{\chi_k}, (k = 1, \ldots, \varkappa)$ ▷ See eq. (4.6)
      for $k = 1, \ldots, \varkappa$ **do**
        if $\psi_k < \phi_{\theta^{(n-1)}}(Y^\circ_{\chi_k})$ **then**
          Set `accepted` ← `False`
    until `accepted`
    Set $Y_{\widetilde{\chi}_k} \leftarrow Y^\circ_{\widetilde{\chi}_k}, (k = 1, \ldots, \varkappa)$ ▷ Accepted $\mathscr{S}$
  Draw $\theta^\circ \sim q(\theta^{(n-1)}, \cdot)$
  Draw $U \sim$ `Unif([0,1])`
  if $U < a(\theta^{(n-1)}, \theta^\circ)$ **then** ▷ See eqs. (4.11) and (4.12)
    Set $\theta^{(n)} \leftarrow \theta^\circ$
  else
    Set $\theta^{(n)} \leftarrow \theta^{(n-1)}$
return $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant measure $\pi(\theta|\mathscr{D})$

Sermaidis et al. (2013) address one more issue: to accelerate mixing of the θ-chain, a correlation between the simulated Poisson points Φ and a value of the parameter θ can sometimes be reduced by defining Φ in a non-centred way. To this end, a non-centred process $\widetilde{\Phi} : \{(\chi_j, \psi_j); j = 1, \ldots\}$ is defined as a unit intensity Poisson point process on an infinite slab $[t_i, t_{i+1}] \times [0, \infty)$ and a function $\widetilde{\Psi}^{[i]}_\theta$, that takes $\widetilde{\Phi}$ to Φ is given by:

$$\widetilde{\Psi}^{[i]}_\theta(\widetilde{\Phi}) := \{(\chi, \psi) \in \widetilde{\Phi} : \psi < l^*_\theta(\Upsilon_i)\}, \quad i = 1, \ldots, K.$$

Notice that a surrogate variable for this parametrisation is given by:

$$\mathscr{S} := \bigcup_{i=0}^{K-1} \left\{ \left\{ Z_{\chi_j^{(i)}} ; j = 1, \ldots \right\} \cup \{l^*(\Upsilon_i)\} \right\},$$

which comprises of an infinite number of points, although, as will be clear from eq. (4.13) below, only a finite number of them ever needs to be revealed. The corresponding density for the parameter θ conditioned on the surrogate variable $\mathscr{S}$

and the observations $\mathscr{D}$ is derived in Sermaidis et al. (2013, Theorem 3):

$$\pi(\theta|\mathscr{S},\mathscr{D}) \propto \pi(\theta)\exp\{A_\theta(\Psi_\theta(Z)_T) - A_\theta(\Psi_\theta(Z)_0) - l_{*\theta}T\} \cdot \prod_{i=1}^{K}\Big\{D_\theta(x_i)\mathcal{N}_{\Delta_i}[\eta_\theta^{-1}(x_{i+1}) - \eta_\theta^{-1}(x_i)] \cdot \prod_{j=1}^{\infty}\Big[1 - \mathbb{1}_{(-\infty, l^*_\theta(\Upsilon_i))}(\psi_j^{(i)})\phi_\theta(\Psi_\theta^{[i]}(Z)_{\chi_j^{(i)}})/l^*_\theta(\Upsilon_i)\Big]\Big\}. \tag{4.13}$$

Notice that only $(\chi,\psi) \in \widetilde{\Phi}$, for which $\psi < l^*_\theta(\Upsilon_i)$ contribute to the value of eq. (4.13) and there is only an almost surely finite number of those points.

Before summarising this doubly non-centred algorithm (non-centred for paths Z and non-centred for the Poisson points $\widetilde{\Phi}$), one more obstacle needs to be overcome. Notice that if Brownian bridges are sampled via localisation construction, then for a new parameter draw θ° it is possible that $l^*_{\theta^\circ}(\Upsilon_i) > l^*_{\theta^{(n-1)}}(\Upsilon_i)$ for some $i \in \{0,\dots,K-1\}$. Since $\mathscr{S}$ is drawn and accepted under the assumption that $\theta = \theta^{(n-1)}$, a Poisson point process $\widetilde{\Phi}$ needs to be sampled only on $[t_i, t_{i+1}] \times [0, l^*_{\theta^{(n-1)}}(\Upsilon_i)]$, $(i = 0,\dots,K-1)$ and Y revealed accordingly for this step. Nonetheless, if the same path Y were accepted under the assumption of $\theta = \theta^\circ$, then, a Poisson point process $\widetilde{\Phi}$ would have needed to be simulated on a strictly larger rectangle $[t_i, t_{i+1}] \times [0, l^*_{\theta^\circ}(\Upsilon_i)]$. Consequently, to evaluate $\pi(\theta^\circ|\mathscr{S},\mathscr{D})$ in eq. (4.13), $\mathscr{S}$ needs to be revealed at additional time points corresponding to the time-positions of a unit intensity Poisson point process on $[t_i, t_{i+1}] \times (l^*_{\theta^{(n-1)}}(\Upsilon_i), l^*_{\theta^\circ}(\Upsilon_i)]$. A naïve way of solving this issue is to sample additional points of $\widetilde{\Phi}$ post-acceptance of $\mathscr{S}$ and reveal the latter at additional time points via techniques from, say, Pollock (2013). However, this approach involves some relatively inefficient computational routines and is inferior to a simple re-arrangement of the simulation steps.

For the rearrangement, θ° is drawn from q first; then, a common upper bound $l^*_{\theta^{(n-1)},\theta^\circ}(\Upsilon_i) := l^*_{\theta^\circ}(\Upsilon_i) \wedge l^*_{\theta^{(n-1)}}(\Upsilon_i)$ is set, a Poisson point process $\widetilde{\Phi}$ is drawn on a rectangle $[t_i, t_{i+1}] \times [0, l^*_{\theta^{(n-1)},\theta^\circ}(\Upsilon_i)]$ and $\mathscr{S}$ revealed at time-points corresponding to all simulated Poisson points. Only those points which fall on $[t_i, t_{i+1}] \times [0, l^*_{\theta^{(n-1)}}(\Upsilon_i)]$ are used in the accept/reject step of the path space rejection sampler. However, now, post-acceptance, $\mathscr{S}$ is already known at all points needed for the evaluation of $\pi(\theta^\circ|\mathscr{S},\mathscr{D})$, even if $l^*_{\theta^\circ}(\Upsilon_i) > l^*_{\theta^{(n-1)}}(\Upsilon_i)$. A complete formulation of this doubly

non-centrally parametrised exact inference algorithm of Sermaidis et al. (2013) is given in algorithm 4.9.

Algorithm 4.9 Doubly non-centred exact inference for diffusion processes

Initialise $\theta^{(0)}$
for $n = 1, \ldots, N$ **do**
  Draw $\theta^\circ \sim q(\theta^{(n-1)}, \cdot)$
  for $i = 0, \ldots, K-1$ **do**
    repeat ▷ Path imputation on $[t_i, t_{i+1}]$ via exact rejection sampling
      Set `accepted` ← `True`
      Draw $\Upsilon^{[i]}$ ▷ Layer information, see section 3.1
      Set $l^*_{\theta^{(n-1)},\theta^\circ}(\Upsilon_i) \leftarrow l^*_{\theta^\circ}(\Upsilon_i) \wedge l^*_{\theta^{(n-1)}}(\Upsilon_i)$
      Draw $\widetilde{\Phi} := \{(\widetilde{\chi}_k, \widetilde{\psi}_k); k = 1, \ldots, \widetilde{\varkappa}\} \sim \mathbb{L}$ on $[t_i, t_{i+1}] \times [0, l^*_{\theta^{(n-1)},\theta^\circ}(\Upsilon_i)]$
      Set $\Phi := \{(\chi_k, \psi_k); k = 1, \ldots, \varkappa\} \leftarrow \{(\chi, \psi) \in \widetilde{\Phi} : \psi < l^*_{\theta^\circ}(\Upsilon_i)\}$
      Draw $W^\circ \sim \mathbb{P}_0(\cdot|\mathscr{D}|_{[t_i,t_{i+1}]} \cap \Upsilon_i)$ at times $\{\widetilde{\chi}_k; k = 1, \ldots, \widetilde{\varkappa}\}$
      Set $Y^\circ_{\chi_k} \leftarrow [\Psi^{[i]}_{\theta^{(n-1)}}(W^\circ)]_{\chi_k}, (k = 1, \ldots, \varkappa)$ ▷ See eq. (4.6)
      for $k = 1, \ldots, \varkappa$ **do**
        if $\psi_k < \phi_{\theta^{(n-1)}}(Y^\circ_{\chi_k})$ **then**
          Set `accepted` ← `False`
    until `accepted`
    Set $Y_{\widetilde{\chi}_k} \leftarrow [\Psi^{[i]}_{\theta^{(n-1)}}(W^\circ)]_{\widetilde{\chi}_k}, (k = 1, \ldots, \widetilde{\varkappa})$ ▷ Reveal Y at all Poisson points $\widetilde{\Phi}$
  Draw $U \sim$ `Unif([0,1])`
  if $U < a(\theta^{(n-1)}, \theta^\circ)$ **then** ▷ See eqs. (4.12) and (4.13)
    Set $\theta^{(n)} \leftarrow \theta^\circ$
  else
    Set $\theta^{(n)} \leftarrow \theta^{(n-1)}$
return $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant measure $\pi(\theta|\mathscr{D})$

4.3 Review of alternative methods

For the setting of exact observations, many frequentist inference methods are based on maximising the log-likelihood function:

$$\log \pi(\mathscr{D}|\theta) = \sum_{i=0}^{K-1} \log\Big[p^{(\theta)}_{t_{i+1}-t_i}(X_{t_i}, X_{t_{i+1}}) \Big]. \tag{4.14}$$

Dacunha-Castelle and Florens-Zmirou (1986) prove that under mild regularity conditions on X, using (4.14) to derive a maximum likelihood estimate (MLE) of θ re-

sults in consistent and asymptotically (as $K \to \infty$) Normal estimators. Of course, as $p_t^{(\theta)}$ is typically intractable, approximations $\widehat{p}_t^{(\theta)}$ to $p_t^{(\theta)}$ must be employed. The simplest—Euler-type—approximation is obtained by fixing the values of the drift and volatility coefficients of X on each $[t_i, t_{i+1})$, $(i = 0, \ldots, K-1)$ (to values taken at the observation times). This results in a process $\widehat{X}$ with piece-wise linear coefficients. The transition densities $\widehat{p}_t^{(\theta)}$ of such a process, evaluated at the observation times, are tractable and can be used as a proxy for $p_t^{(\theta)}$. This approach is analysed by Florens-Zmirou (1989) and Yoshida (1992). Similarly motivated idea has been proposed by Shoji (1998) and Shoji and Ozaki (1998a); however, there, an improved approximation $\widehat{X}$ to the true process X is investigated.

Potentially more accurate estimators to $p_t^{(\theta)}$, based on solving Kolmogorov forward (or backward) equations—which $p_t^{(\theta)}$ is known to solve under mild regularity conditions on X (Karatzas and Shreve, 1998a, p. 368–369)—have been studied for instance in Poulsen (1999). Standard numerical techniques for partial differential equations can be used to derive pointwise estimators of $p_t^{(\theta)}$. Unfortunately, such approach is often computationally prohibitively expensive.

Another type of approximations to $p_t^{(\theta)}$ was suggested by Aït-Sahalia (2002) (assuming A6 and A7) and Aït-Sahalia (2008) (for more general diffusions). The authors propose to expand $p_t^{(\theta)}$ around Gaussian densities using Hermite polynomials and define $\widehat{p}_t^{(\theta)}$ by truncating these expansions after including a sufficiently large number of terms.

Pedersen (1995) exploit Chapman-Kolmogorov equations to write a transition density $p_t^{(\theta)}$ as the convolution of $p_{s_{j+1}-s_j}^{(\theta)}$, $j = 1, \ldots, m-1$ defined over sub-intervals $[s_j, s_{j+1}]$ of t, $(0 = s_0 < \cdots < s_m = t)$. This leads to a representation of the transition density $p_t^{(\theta)}$ as an expectation of certain (Gaussian) function over the path measure of an unconditioned diffusion. Consequently, a Monte Carlo estimator (based on repeated simulations of unconditioned diffusions) can be used in place of $\widehat{p}_t^{(\theta)}$. An additional importance sampling step, introduced in Durham and Gallant (2002), allows to further improve the quality of $\widehat{p}_t^{(\theta)}$.

Dacunha-Castelle and Florens-Zmirou (1986) find an important representation of $p_t^{(\theta)}$ as an expectation with respect to a (tractable) Wiener measure (which can be derived from eqs. (2.8) and (2.10) by taking expectation on both sides of eq. (2.10)

with respect to a measure $\mathbb{P}_0(\cdot|\mathcal{Z})$ and realising that the left hand side is equal to 1).

$$p_t^{(\theta)}(x_0, x_t) = |\det(\sigma_\theta)|^{-1} \mathcal{N}_t(y_t - y_0) \mathbb{E}_{\mathbb{P}_0}\left[\exp\left\{ A_\theta(Y_t) - A_\theta(Y_0) - \int_0^t \varphi_\theta(Y_s)\,\mathrm{d}s \right\} \middle| Y_t = y_t \right], \tag{4.15}$$

where as usual $y_u := \eta_\theta(x_u)$, $u \in \{0, t\}$. Nicolau (2002) use representation in eq. (4.15) to define $\widehat{p}_t^{(\theta)}$ as a suitable Monte Carlo estimator.

Beskos et al. (2006, 2008, 2009) and Fearnhead et al. (2008) refine the idea of Nicolau (2002), by proposing more elaborate Monte Carlo estimators based on eq. (4.15). These can be equipped with a desirable property of unbiasedness and some may even be restricted to be positive and bounded. Monte Carlo Expectation-Maximisation and a number of other techniques can then be used to find MLEs for such estimators (Beskos et al., 2006).

Another successful strand of frequentist techniques are based on defining suitable estimating functions $E_K(\theta) := E_K(\theta, X_1, \ldots, X_K)$ (see Bibby et al. (2010) or Kessler et al. (2012) for a comprehensive review), ideally equipped with a property:

$$\mathbb{E}[E_K(\widehat{\theta})] = 0 \iff \widehat{\theta} \text{ was used to generate the data.}$$

$\widehat{\theta}$ is sought by finding the roots of $\theta \mapsto \mathbb{E}[E_K(\theta)]$. The *if and only if* statement is sometimes too difficult to satisfy in practice and therefore less efficient estimators are often admissible.

A natural environment for Bayesian techniques is that of data-augmentation and thus many "alternative" methods to the ones considered in this thesis can be obtained by simply changing a sampler of the conditioned diffusions (see section 3.4 for examples of those). Additionally, unbiased and positive estimators of $p_t^{(\theta)}$ due to Fearnhead et al. (2008) (mentioned above) can be used to define pseudo-marginal samplers (Andrieu and Roberts, 2009).

Some methods above—such as most schemes based on data-augmentation, as well as techniques relying on approximations via Gaussian densities (say, Durham and Gallant (2002))—are particularly easy to extend to the setting of partially observed diffusions with noise (see respective sources). Others—and especially those relying on numerical approximations to the transition densities—are restricted mainly to the exact observational setting. However, only a handful of the approaches above have seen extensions to other observational regimes.

For instance, observations of the first passage times of scalar diffusions (used in so-called leaky integrate-and-fire models) have been treated almost exclusively in the context of frequentist literature and within a limited range of diffusion processes (see Lansky and Ditlevsen (2008) for a review). They are mainly restricted to methods based on Fortet's equation (Fortet, 1943) (relating transition density of the process to its first passage time density) or on the recursive formulas for the moments of the first passage time distribution (Siegert, 1951), and include numerical approximations to the first passage time densities, maximum likelihood estimators or employments of methods of moments to find other estimators (Ricciardi and Sato, 1988; Ricciardi et al., 1999; Ditlevsen and Lansky, 2005, 2006, 2007; Ditlevsen and Ditlevsen, 2008). A single Bayesian approach has been put forth in Iolov et al. (2017) and it is based on solving partial differential equations (arising from combining Fokker-Planck equations with a problem of maximising mutual information between parameters and the data). To the best of my knowledge no results are known on the more elaborate, composite first passage time bridges presented in the Introduction, despite such observational setting being more faithful to the mechanisms underlying phenomena studied in neuroscience (Leon et al., 2018).

Methods discussed thus far are suitable for the parametric setting (i.e. when the functional form of b_θ and σ_θ are known up to a finite-dimensional parameter θ). The non-parametric setting (i.e. when θ is infinite-dimensional and the inference is done directly on the space of functions), has also been studied extensively in the frequentist context, however Bayesian literature is lacking. The former is dominated by kernel-type estimators. Examples include those of Banon (1978); Stanton (1997), locally linear smoothers with adaptive bandwidth (Spokoiny, 2000), and estimators derived via penalised likelihood (Comte et al., 2007). Showing consistency and contraction rates of the estimators is, however, non-trivial (Dalalyan and Kutoyants, 2002; Gobet et al., 2004; Tuan, 1981; van Zanten, 2001). van Zanten (2013) provide a review of the short literature on the Bayesian non-parametric inference methods. To the best of my knowledge, in the setting of low-frequency observations only methods of inference of the drift coefficient have been described to date. In Papaspiliopoulos et al. (2012) the unknown drift function is equipped with a

prior measure in function space, which is taken to be Gaussian with mean 0 and covariance defined by a certain differential operator. A data-augmentation scheme (Roberts and Stramer, 2001) is then used to compute an MCMC approximation to the posterior. Consistency results for this setting are shown by Pokern et al. (2013), and improved contraction rates are derived by van Waaij and van Zanten (2016). van der Meulen et al. (2014) proposed an algorithmic modification to the procedure of Papaspiliopoulos et al. (2012), using a different basis expansion for the drift function and employing random truncation of this expansion (with the truncation point equipped with a prior and explored with a reversible jump step). Contraction rates for this approach were derived in van der Meulen et al. (2018). Finally, Gugushvili et al. (2018) recently proposed a Bayesian method of inference of the diffusion coefficient for densely observed processes.

4.4 Commentary

Bayesian inference is not the only application for simulation methods of conditioned diffusions. Another extensive area of research where these techniques find use is in estimating rare event probabilities (Cérou et al., 2012; Vanden-Eijnden and Weare, 2012). Example 2.4.1 had a flavour of many of the problems treated in this field. Another application comes from the literature on signal processing and it has to do with non-linear filtering (Jazwinski, 2007). Just like in sequential Monte Carlo algorithms, the signal is known at time 0: X_0, and its noisy observation is made at time T: $LX_T + \xi$. The interest lies in estimating the marginal distribution of a signal at time T, given $(X_0, LX_T + \xi)$. A third application is to estimating expectations over path functionals. These last problems are of particular relevance for pricing financial options (Fournié et al., 2001).

PART II

Extensions

5

Reducing computational cost with blocking

This chapter is based on joint work with Paul A. Jenkins, Murray Pollock and Gareth O. Roberts.

Exact rejection sampling on a path space due to Beskos and Roberts (2005); Beskos et al. (2006, 2008) (summarised in section 3.1) has one distinguishing feature that makes it stand out among other diffusion bridge simulation algorithms—it is devoid of any discretisation errors. Consequently, as shown in section 4.2 it can be applied to Bayesian inference for diffusion processes and it is guaranteed that no bias (beyond that of a burn-in and finiteness of the Markov chain) enters the resulting inference algorithm. Some competing methods—such as guided proposals and simple diffusion bridges—allow to exert direct control over introduced bias, making it possible to shrink it to arbitrarily small levels by increasing the density of the grid at which paths are imputed; however, this comes at an expense of increased computational cost. If the bias tolerance is sufficiently low, then the incurred costs might be prohibitively large.

Nonetheless, exact rejection sampling on a path space is not without its drawbacks—I discuss those in section 2.3.4. One of them is unfavourable scaling of its computational cost with the duration of the bridges. More precisely, as I showed in section 3.1.4, the computational cost of obtaining any diffusion bridge sample increases exponentially with T. Consequently, exact rejection sampling on a path space is not well suited for the problem of Bayesian inference for sparsely observed diffusions. This shortcoming is a motivation for a detailed analysis of the *blocking* technique in this chapter.

Heuristically, blocking breaks down the sampling of an entire trajectory into independent problems of sampling paths on shorter time segments, effectively transforming a rejection sampler on a path space into a Gibbs sampler on a path space. The methodology does not limit the class of diffusions to which rejection sampling on a path space can be applied to, and in this chapter I show that it leads to scaling

of the computational cost that is only polynomial in T (which in some instances can be shown to be slightly in excess of cubic).

The blocking technique goes back at least to Shephard and Pitt (1997), where it was used in the context of time-series analysis. In the context of simulating diffusion processes this technique has been employed by a number of authors: Chib et al. (2004); Kalogeropoulos (2007); Kalogeropoulos et al. (2010); Golightly and Wilkinson (2008); Fuchs (2013); van der Meulen and Schauer (2018) and many others, with Stramer and Roberts (2007) employing it even to NLCAR models (continuous-time non-linear autoregressive models). Furthermore, some authors (such as Sermaidis et al. (2013)), mentioned blocking in passing exactly in the setting of rejection sampling on a path space as a possible extension. Nonetheless, despite the apparent popularity of this technique no formal quantification has ever been provided for the reduction in the computational cost arising from the use of blocking.

It turns out that such quantification can be made reasonable precise under the choices of simple diffusion processes: Brownian bridges and the Ornstein-Uhlenbeck bridges. The conclusions drawn from these two examples are similar: in both cases the blocking-imbued rejection sampling on a path space has computational cost that scales at a rate that can be bounded by a quartic function of the duration of the bridges (in fact, a relation only slightly larger than cubic is needed). Despite the results being limited to the two special diffusions above, the numerical examples given in section 5.4 indicate much broader applicability of those conclusions and include empirical evidence supporting extension to substantially more non-linear diffusion processes.

5.1 Blocking technique

As usual, let $\mathbb{P}_b$ denote the unconditioned diffusion law induced by the SDE (1.1) and consider a problem of drawing paths from the conditioned law $\mathbb{P}_b(\cdot|\mathcal{Z})$, where for simplicity of exposition I take $\mathcal{Z} := X_T$. Rejection sampling on a path space proceeds by transforming $\mathbb{P}_b(\cdot|\mathcal{Z})$ to $\mathbb{P}_\alpha(\cdot|\mathcal{Z})$ (induced by (2.6)) and proposing paths from $\mathbb{P}_0(\cdot|\mathcal{Z})$ until the first acceptance (see section 2.3 for details).

To modify this algorithm with a blocking technique, first, define a set of *stochastic knots* (which I refer to as *knots* for short):

$$\mathscr{K}(\omega) := \{X_{k_1}(\omega), \ldots, X_{k_m}(\omega)\},$$

spread across the time domain: $0 < k_1 < \cdots < k_m < T$. Now, divide the knots into $\mathbb{k}$ disjoint subsets:

$$\mathscr{K}_i(\omega) := \{X_{r_{ij}}(\omega);\, j = 1, \ldots, m_i\}, \quad i = 1, \ldots, \mathbb{k},$$
$$(\text{where } \cup_i \mathscr{K}_i = \mathscr{K} \text{ and } \mathscr{K}_i \cap \mathscr{K}_j = \varnothing \text{ if } i \neq j).$$

For notational convenience, define also the set of all knots which do not belong to $\mathscr{K}_i$:

$$\mathscr{K}_{-i}(\omega) := \bigcup_{j \neq i} \mathscr{K}_j(\omega), \quad i = 1, \ldots, \mathbb{k},$$

and the expanded sets of conditioned-on random variables:

$$\mathscr{Z}_{-i}(\omega) := \mathscr{Z} \cup \mathscr{K}_{-i} \cup \{X_0\} := \{X_{e_j^{(i)}}(\omega);\, j = 0, \ldots, (m - m_i + 1)\}, \tag{5.1}$$

(where $e_j^{(i)}$ are simply new index labels) so that in the case of $\mathscr{Z} := X_T$, $\mathscr{Z}_{-i}(\omega)$ consist of an initial point, an end point and all stochastic knots which do not belong to $\mathscr{K}_i$. Define also the convenience function ι, which takes a set of knots (or expanded set of conditioned-on random variables) and returns the times at which the knots are anchored, so, for instance:

$$\iota(\mathscr{K}_i) := \{r_{ij};\, j = 1, \ldots, m_i\}, \quad \text{or} \quad \iota(\mathscr{Z}_{-i}) := \{e_j^{(i)};\, j = 0, \ldots, (m - m_i + 1)\}.$$

Finally, define

$$\mathscr{E}_i := \left\{ (e_j^{(i)}, e_{j+1}^{(i)}) \,\middle|\, \exists r \in \iota(\mathscr{K}_i) \text{ s.t. } r \in [e_j^{(i)}, e_{j+1}^{(i)}] \right\}_{j=0}^{m-m_i},$$

to be only those intervals in-between the end-points or knots from $\mathscr{Z}_{-i}$, which contain at least one knot belonging to $\mathscr{K}_i$. The path segments $X|_{\mathscr{E}_i}$, obtained through restricting X to $\mathscr{E}_i$, are termed *blocks*. Knots and blocks are illustrated graphically in fig. 5.1.

A blocked rejection sampler on a path space is an MCMC algorithm (or, more precisely, a Gibbs sampler), which instead of updating path X in full by sampling

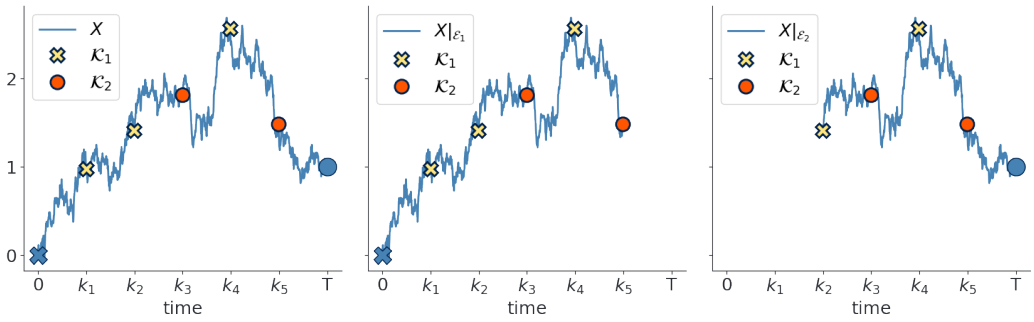


Figure 5.1: Illustration of knots and blocks. On left-most plot an entire trajectory of X is drawn. Two sets of knots are given ($\Bbbk = 2$). $\mathscr{K}_1 = \{X_{k_1}, X_{k_2}, X_{k_4}\}$ are marked with yellow crosses and $\mathscr{K}_2 = \{X_{k_3}, X_{k_5}\}$ are marked with orange dots. In the middle plot, only blocks belonging to $X|_{\mathcal{E}_1}$ are drawn, whereas in the right-most plot only blocks belonging to $X|_{\mathcal{E}_2}$ are given.

from $\mathbb{P}_0(\cdot|\mathcal{Z})$, updates blocks $X|_{\mathcal{E}_i}$, $(i = 1, \ldots, \Bbbk)$ sequentially and multiple times. The steps are given explicitly in the algorithm 5.1 below.

Algorithm 5.1 Blocked rejection sampler on a path space

1: Initialise X
2: **for** $n = 1, \ldots, N$ **do**
3: **for** $i = 1, \ldots, \Bbbk$ **do** ▷ One full blocking sweep
4: Draw $I \sim q_I(i, \cdot)$ ▷ $q_I(i, \cdot)$ is a user-chosen pmf, see below
5: Draw $X|_{\mathcal{E}_I} \sim \mathbb{P}_b|_{\mathcal{E}_I}(\cdot|\mathcal{Z}_{-I})$ ▷ Using PSRS, see below

Each step of the Gibbs sweep starts in line 4 of algorithm 5.1, where it is decided which block is to be updated. Different transition kernels $q_I(i, \cdot)$ give raise to different updating strategies—below, I define some of them. After the block is picked, it is then updated in line 5 of algorithm 5.1 by drawing from $X|_{\mathcal{E}_I} \sim \mathbb{P}_b|_{\mathcal{E}_I}(\cdot|\mathcal{Z}_{-I})$. By the Markov property the law $\mathbb{P}_b|_{\mathcal{E}_I}(\cdot|\mathcal{Z}_{-I})$ factorises at the times $t \in \iota(\mathscr{K}_{-I})$ and thus updates of $X|_{[e_j^{(i)}, e_{j+1}^{(i)}]}$ on each sub-interval $[e_j^{(i)}, e_{j+1}^{(i)}]$ can be done independently from one another using rejection sampling on a path space. An entire, single sweep is illustrated graphically in fig. 5.2.

Definition 5.1.1. Let $\Bbbk = 2$. $\mathscr{K}_1$ and $\mathscr{K}_2$ are termed *interlaced* if whenever $a, c \in \iota(\mathscr{K}_i)$, with $a < c$, then there exists $b \in \iota(\mathscr{K}_{(i \bmod 2)+1})$ s.t. $a < b < c$, $i = 1, 2$.

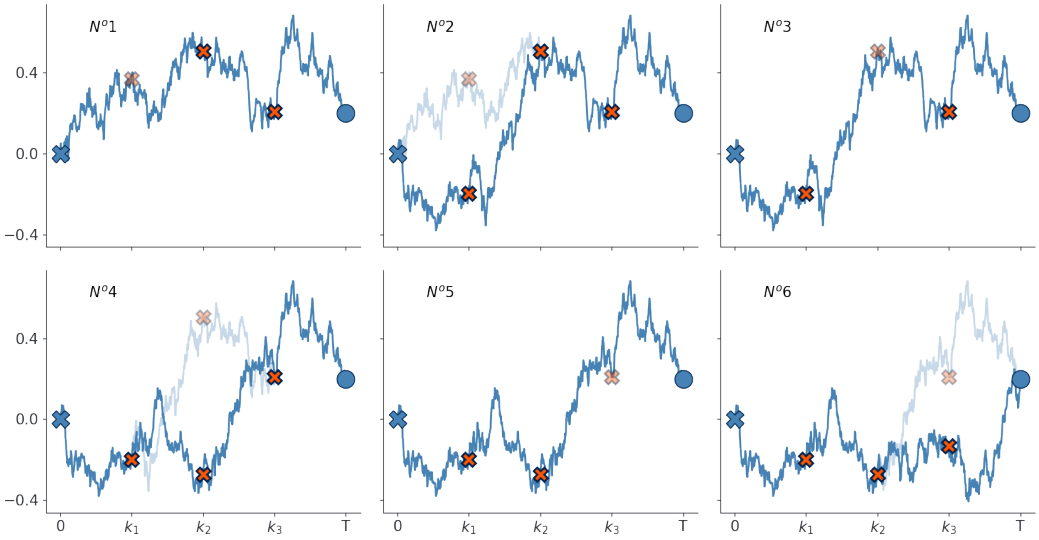


Figure 5.2: Single sweep (done in the lexicographic order) of a blocked rejection sampler on a path space. $N°1$: X needs to be sampled on the segment $\mathscr{E}_1 = (0, k_2)$, so the knot X_{k_1} is temporarily forgotten. $N°2$: the update is performed by sampling from the law $\mathbb{P}_b|_{\mathscr{E}_1}(\cdot|\mathcal{Z}_{-1})$. This leads to the knot X_{k_1} taking a new value. $N°3$: X needs to be sampled on the segment $\mathscr{E}_1 = (k_1, k_3)$, so the knot X_{k_2} is temporarily forgotten. $N°4$: the update is performed by sampling from the law $\mathbb{P}_b|_{\mathscr{E}_2}(\cdot|\mathcal{Z}_{-2})$. This leads to the knot X_{k_2} taking a new value. $N°5$: X needs to be sampled on the segment $\mathscr{E}_1 = (k_2, T)$, so the knot X_{k_3} is temporarily forgotten. $N°6$: the update is performed by sampling from the law $\mathbb{P}_b|_{\mathscr{E}_3}(\cdot|\mathcal{Z}_{-3})$. This leads to the knot X_{k_3} taking a new value. The cycle is then repeated.

Definition 5.1.2. (CBUS) Let $\mathbb{k} = 2$, $\mathscr{K}_1$ and $\mathscr{K}_2$ be interlaced and $q_I(i,j) := \mathbb{1}_{\{i\}}(j)$. Then, I refer to algorithm 5.1 as a chequerboard blocking update scheme—CBUS for short.

Definition 5.1.3. (LBUS) Let $\mathbb{k} = m$, $\mathscr{K}_i(\omega) := \{X_{k_i}(\omega)\}$, $i = 1, \ldots, m$ and $q_I(i,j) := \mathbb{1}_{\{i\}}(j)$. Then, I refer to algorithm 5.1 as a lexicographic blocking update scheme—LBUS for short.

Definition 5.1.4. (RBUS) Let $\mathbb{k} = m$, $\mathscr{K}_i(\omega) := \{X_{k_i}(\omega)\}$, $i = 1, \ldots, m$ and $q_I(i,j) := \frac{1}{m}\mathbb{1}_{\{1,\ldots,m\}}(j)$. Then, I refer to algorithm 5.1 as a random blocking update scheme—RBUS for short.

5.2 Quantifying computational cost

As discussed in section 3.1.4 the expected computational cost of a single call to path space rejection sampler is:[1]

$$\mathscr{R}_{PSRS}(T) = \Theta(T\exp\{cT\}), \quad \text{as } T \to \infty, \tag{5.2}$$

and each such call outputs an independent draw from the target law. For simplicity I assume that the knots are set on an equidistant time-grid $k_0 := 0 < k_1 < \cdots < k_m < T =: k_{m+1}$, with:

$$\Delta := \Delta(T,m) := \frac{T}{m+1} = k_{i+1} - k_i, \qquad i = 0,\ldots,m,$$

so that `CBUS`, `LBUS` and `RBUS` always update segments of X on intervals of length 2Δ. In particular, each call to path space rejection sampler made under any of the schemes above involves updates on intervals of length 2Δ, and not T. This means that the cost of a single sweep (lines 3–5 of algorithm 5.1) under any of the three updating strategies is given by:

$$\mathscr{R}_{sweep}(T,m) = \Theta(m\mathscr{R}_{PSRS}(2\Delta)) = \Theta(T\exp\{2c\Delta\}), \quad \text{as } T \to \infty. \tag{5.3}$$

The exponential explosion in the computational expense in (5.2) can therefore be combated by increasing m (the number of knots). However, this remedy has its price—the larger the m, the more correlated the samples of a single Gibbs sweep become. Since it is only fair to compare rejection sampling on a path space with the algorithm that outputs "almost independent" samples from the target law, taking larger m means that a greater number of steps N needs to be taken before the blocked sampler "forgets" the initial state of the path. Therefore, it is apparent that blocking involves a non-trivial trade-off of costs.

To formally quantify this trade-off, a rigorous definition of "forgetting" needs to be proposed. To this end, first notice that instead of analysing the chain targeting the law of paths $X|(X_0,X_T)$: $\mathbb{P}_b(\cdot|\mathcal{Z})$, it is sufficient to consider the chain targeting the law $\mathbb{P}^*_b(\cdot|\mathcal{Z})$, defined as the marginal law of the vector

$$\mathscr{G} := (X_{k_1},\ldots,X_{k_m})|(X_0,X_T) = \mathscr{K}|(X_0,X_T). \tag{5.4}$$

[1] In this section $\mathcal{O}(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ and $o(\cdot)$ take the meaning defined by the standard, Bachmann–Landau notation.

To see this, notice that conditionally on the knots $\mathscr{K}$ being distributed exactly from $\mathbb{P}_b^*(\cdot|\mathcal{Z})$, a path X returned after a single Gibbs sweep of, say, `CBUS` is distributed exactly from $\mathbb{P}_b(\cdot|\mathcal{Z})$. Next, define: P^n to be the n-step transition kernel of algorithm 5.1, with a single step P corresponding to an entire Gibbs sweep (lines 3–5). For conciseness of notation let $\pi := \mathbb{P}_b^*(\cdot|\mathcal{Z})$. Then, I use the following statement as a proxy for "forgetting" the initial state.

Definition 5.2.1. A Markov chain with the transition kernel P, an invariant density π and started from $\mathscr{G}^{(0)}$ is said to output a sample after n steps that is *independent at an ϵ-tolerance level* if

$$||P^n(\mathscr{G}^{(0)},\cdot) - \pi(\cdot)||_{TV} \leq \epsilon. \tag{5.5}$$

To justify the use of definition 5.2.1, it can be shown that if eq. (5.5) is satisfied then a random variable $Z^{(\mathscr{G}^{(0)})} \sim P^n(\mathscr{G}^{(0)},\cdot)$ admits the following decomposition:

$$Z^{(\mathscr{G}^{(0)})} = (1-B)Z + BH,$$

where $Z \sim \pi(\cdot)$, $B \sim$ `Bernoulli(`ϵ`)` and H is some independent random variable (see Huber (2016, Fact 3.3) for a proof). Therefore, with probability $(1-\epsilon)$ the value taken after the n^{th} step of a Markov chain is distributed exactly according to the target law—in particular, it is independent from the initial state $\mathscr{G}^{(0)}$.

Denote with $N := N(T, m, \epsilon, \mathscr{G}^{(0)})$ the total number of steps of the Gibbs sampler in the algorithm 5.1 for which output is independent at an ϵ-tolerance level. Once N is found, it can be used in conjunction with the estimate for the computational cost of a single sweep from eq. (5.3) so as to derive the asymptotic (as $T, m \to \infty$) computational cost of blocked rejection sampling on a path space:

$$\mathscr{R}_{blocking}(T, m, \epsilon, \mathscr{G}^{(0)}) = \mathcal{O}\left(N(T, m, \epsilon, \mathscr{G}^{(0)}) T \exp\{2c\Delta(T, m)\}\right), \quad \text{as } T, m \to \infty.$$

To derive $N(T, m, \epsilon, \mathscr{G}^{(0)})$ I will resort to results on convergence rates of Markov chains. A natural metric for measuring convergence rate of a Gibbs sampler is $\mathcal{L}^2$-norm (Amit, 1991). More precisely, define

Definition 5.2.2. (Roberts and Sahu, 1997) The convergence rate ρ of a Markov chain with the transition kernel P and an invariant density π is defined as the minimum number for which for all square π-integrable functions f, and for all $r > \rho$

$$||P^n f(\mathscr{G}^{(0)}) - \pi(f)||_{\mathcal{L}^2(\pi)} := \mathbb{E}_\pi\left[\{P^n f(\mathscr{G}^{(0)}) - \pi(f)\}^2\right] \leq V_f(\mathscr{G}^{(0)}) r^n,$$

where $P^n f(\mathscr{G}^{(0)}) := \mathbb{E}_\pi[f(\mathscr{G}^{(n)})|\mathscr{G}^{(0)}]$, $\pi(f) := \mathbb{E}_\pi[f(\mathscr{G})]$ and V_f is some function of $\mathscr{G}^{(0)}$, allowed to depend on f.

I will derive the asymptotic (as $T, m \to \infty$) convergence rate of the Markov chain defined by algorithm 5.1 as a function of $\mathscr{G}^{(0)}$, ϵ, T and m, under the assumption that $\mathbb{P}_b$ is either the law of Brownian motion or the law of the Ornstein-Uhlenbeck process. I will then show how to translate the derived statement about $\mathcal{L}^2$ convergence rate into the corresponding statement regarding the total variation distance of the form of eq. (5.5). Expression for $N(T, m, \epsilon, \mathscr{G}^{(0)})$ will then follow.

5.2.1 $\mathcal{L}^2$ convergence rate

Assumption A15. *The target law $\mathbb{P}_b$ is such that $\mathscr{G}$ is a Gaussian process.*

Assume throughout that A15 above holds. Restriction to Gaussian processes is the key technical simplification that renders many of the calculations tractable, making it possible to derive closed form expressions for the asymptotic convergence rates.

A single Gibbs step of algorithm 5.1 (lines 4 and 5) has a tractable transition density and so does an entire sweep $\mathscr{G}^{(n)} \to \mathscr{G}^{(n+1)}$ (lines 3–5 of algorithm 5.1). Denote by Σ the covariance matrix of $\mathscr{G}$. By Roberts and Sahu (1997, Lemma 1), the chain $\{\mathscr{G}^{(n)}; n = 0, \dots\}$ has the Normal transition density with mean and covariance matrix given respectively by:

$$\mathbb{E}[\mathscr{G}^{(n+1)}|\mathscr{G}^{(n)}] = B\mathscr{G}^{(n)} + b, \quad \mathbb{V}ar[\mathscr{G}^{(n+1)}|\mathscr{G}^{(n)}] = \Sigma - B\Sigma B^T, \tag{5.6}$$

where $B \in \mathbb{R}^{m \times m}$ and $b \in \mathbb{R}^m$. Both B and b admit closed form expressions and can be computed using techniques from Roberts and Sahu (1997, §2.2). I omit the details as, in the setting of this chapter, neither object is necessary for finding the

convergence rate ρ. Denote with $\Lambda := \Sigma^{-1}$ the precision matrix of $\mathscr{G}$, and define the matrix A via:

$$A := I - \text{diag}\{\Lambda_{11}^{-1}, \ldots, \Lambda_{mm}^{-1}\}\Lambda.$$

Then, the following holds true:

Lemma 5.2.1. (Roberts and Sahu, 1997) Under A15, the $\mathcal{L}^2$ convergence rate of a blocked rejection sampler on a path space is given by a spectral radius of the matrix B: $\rho_{\texttt{spec}}(B)$. In particular, under CBUS, LBUS and RBUS respectively it can be shown to be equal to:

$$\rho_{\texttt{CBUS}} = \rho_{\texttt{LBUS}} = \lambda_{\texttt{max}}^2(A), \quad \rho_{\texttt{RBUS}} = \left[\frac{m-1+\lambda_{\texttt{max}}(A)}{m}\right]^m,$$

where $\lambda_{\texttt{max}}(A)$ denotes the maximum eigenvalue of the matrix A.

The entries of the precision matrix Λ share a close connection with partial correlations, rendering A simple to compute (it is given in eq. (5.17)). This in turn makes it possible to derive $\lambda_{\texttt{max}}(A)$:

Theorem 5.2.1. Under A15:

$$\lambda_{\texttt{max}}(A) = 2|c(\Delta)|\cos\left(\frac{\pi}{m+1}\right),$$

where $c(\Delta) := \mathbb{C}orr(X_\Delta, X_{2\Delta}|X_0, X_{3\Delta})$. In particular, the $\mathcal{L}^2$ convergence rate of a blocked rejection sampler on a path space under CBUS, LBUS and RBUS respectively is given by:

$$\rho_{\texttt{CBUS}} = \rho_{\texttt{LBUS}} = 4c^2(\Delta)\cos^2\left(\frac{\pi}{m+1}\right), \quad \rho_{\texttt{RBUS}} = \left[\frac{m-1+2|c(\Delta)|\cos\left(\frac{\pi}{m+1}\right)}{m}\right]^m.$$

It is now possible to specialise to particular choices of $\mathbb{P}_b$. Suppose first that $\mathbb{P}_b$ is the law of Brownian motion. Then, $\mathscr{G}$ is a Gaussian process with covariance matrix Σ, whose $(i,j)^{th}$ entry is given by:

$$\Sigma_{i,j} = \frac{(k_i \wedge k_j - k_0)(k_{m+1} - k_i \vee k_j)}{k_{m+1} - k_0},$$

and thus the following holds.

Corollary 5.2.1. If $\mathbb{P}_b$ is the law of a Brownian motion, then the $\mathcal{L}^2$ convergence rate of a blocked rejection sampler on a path space under `CBUS`, `LBUS` and `RBUS` respectively is given by:

$$\rho_{\texttt{CBUS}} = \rho_{\texttt{LBUS}} = \cos^2\left(\frac{\pi}{m+1}\right), \quad \rho_{\texttt{RBUS}} = \left[\frac{m-1+\cos\left(\frac{\pi}{m+1}\right)}{m}\right]^m.$$

In particular, as $m \to \infty$

$$\rho_{\texttt{CBUS}} = \rho_{\texttt{LBUS}} = 1-\left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4}), \quad \rho_{\texttt{RBUS}} = 1-\left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4}). \tag{5.7}$$

On the other hand, if $\mathbb{P}_b$ is taken to be the law of the Ornstein-Uhlenbeck process (which without loss of generality can be centred at 0):

$$\mathrm{d}X_t = -\theta X_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \tag{5.8}$$

then $\mathcal{G}$ is again a Gaussian process, but computation of Σ is more cumbersome. Its form follows from:

$$\mathbb{C}ov\left[\begin{pmatrix} Y_s \\ Y_t \end{pmatrix}\middle| Y_0, Y_T\right] = \frac{\sigma^2}{\theta}\begin{pmatrix} e^{-\theta s}\sinh(\theta s) & e^{-\theta t}\sinh(\theta s) \\ e^{-\theta t}\sinh(\theta s) & e^{-\theta t}\sinh(\theta t)\end{pmatrix} - \frac{\sigma^2}{\theta}\begin{pmatrix} e^{-\theta T}\frac{\sinh^2(\theta s)}{\sinh(\theta T)} & e^{-\theta T}\frac{\sinh(\theta s)\sinh(\theta t)}{\sinh(\theta T)} \\ e^{-\theta T}\frac{\sinh(\theta s)\sinh(\theta t)}{\sinh(\theta T)} & e^{-\theta T}\frac{\sinh^2(\theta t)}{\sinh(\theta T)}\end{pmatrix}, \quad 0 < s < t < T. \tag{5.9}$$

It is now possible to establish the following result.

Corollary 5.2.2. If $\mathbb{P}_b$ is the law of the Ornstein-Uhlenbeck process (5.8), then the correlation $c(\Delta)$ is given by:

$$c(\Delta) = \frac{e^{-2\theta\Delta}\sinh(\theta\Delta) - e^{-3\theta\Delta}\frac{\sinh(\theta\Delta)\sinh(2\theta\Delta)}{\sinh(3\theta\Delta)}}{\sqrt{\left(e^{-\theta\Delta}\sinh(\theta\Delta) - e^{-3\theta\Delta}\frac{\sinh^2(\theta\Delta)}{\sinh(3\theta\Delta)}\right)\left(e^{-2\theta\Delta}\sinh(2\theta\Delta) - e^{-3\theta\Delta}\frac{\sinh^2(2\theta\Delta)}{\sinh(3\theta\Delta)}\right)}}. \tag{5.10}$$

In particular, as $m \to \infty$ and for a fixed T or $T = o(m)$

$$\rho_{\texttt{CBUS}} = \rho_{\texttt{LBUS}} = 1-\left(\frac{a_1}{m+1}\right)^2 + \mathcal{O}(m^{-4}), \quad \rho_{\texttt{RBUS}} = 1-\left(\frac{a_2}{m+1}\right)^2 + \mathcal{O}(m^{-4}),$$

for some constants a_1, a_2.

5.2.2 Relating Total Variation and $\mathcal{L}^2$ distance

$\mathcal{L}^2$ bounds of the form found in definition 5.2.2 are well known to be closely related to statements about the total variation distance of the kind found in definition 5.2.1. By following the methods from (Roberts and Rosenthal, 1997; Papaspiliopoulos et al., prep) I can relate the $\mathcal{L}^2$ convergence rate of the Markov chain $\{\mathcal{G}^{(n)}, n = 0, \dots\}$ to bounds on its total variation distance in the following way:

Theorem 5.2.2. Under the assumption A15 and the assumption that the transition kernel P with invariant density π and $\mathcal{L}^2$ convergence rate $\rho := \rho(m,T)$ is reversible, there exists a function V such that for all $\delta > 0$

$$||P^n(\mathcal{G}^{(0)},\cdot) - \pi(\cdot)||_{TV} \leq V(\mathcal{G}^{(0)},m,T)(\rho(m,T)+\delta)^n. \tag{5.11}$$

It should be noted that out of the three updating schemes presented above only RBUS induces a reversible Markov chain, so in view of the prerequisites of theorem 5.2.2 any further conclusions can be rigorously applied only under the assumption that RBUS was chosen. Nonetheless, it is straightforward to modify both CBUS and LBUS so that the resulting chains are reversible. Indeed, notice that if the regular update of CBUS or LBUS is immediately followed by the same update but in a reversed order, then the resulting chain (consisting of two sweeps) is reversible. Deriving the $\mathcal{L}^2$ convergence rate for such modified chains is possible, but it requires involved computations of matrices and eigenvalues related to matrix B and in view of the same asymptotic results for CBUS, LBUS and RBUS and simultaneous possibility to use RBUS rigorously, derivations of the convergence rates for the modified CBUS and LBUS are omitted.

Conjecture 5.2.1. Dependence of V on m and T (for large enough m and T) is sub-exponential, i.e. $V(\mathcal{G}^{(0)},m,T) \leq c_1(1+m)^{c_2}(1+T)^{c_3} f(\mathcal{G}^{(0)})$ for some function f and some constants $c_1, c_2, c_3 > 0$.

Conjecture 5.2.1 is not going to be possible to be verified in practice, as there is no known method for finding function V explicitly (except for direct computations of total variation, which quickly become overwhelming). Despite that, it is reasonable to believe it to be true. For instance, in the case of $\mathbb{P}_b$ being given by the law of Brownian motion, the matrix of partial correlations of $\mathcal{G}$ (computed

in part in (5.18)) turns out to be entirely independent from T and thus T should have little influence over the Markov chain $\{\mathscr{G}^{(n)}; n = 1,\dots\}$. To investigate the dependence on m, notice from eq. (5.6) that $\{\mathscr{G}^{(n)}; n = 1,\dots\}$ defines an `AR(1)` process (Roberts and Sahu, 1997), and thus all eigenvalues of matrix B must be smaller in magnitude than unity ($\rho_{\text{spec}}(B) < 1$). Meyn and Tweedie (2012, §16.5.1) show that for the appropriate multidimensional random coefficient autoregressive models (defined by the recurrence $X_{n+1} = (B + \Gamma_n)X_n + W_n$ for a sequence of random matrices Γ_n) the Foster-Lyapunov drift condition (Meyn and Tweedie, 2012, p. 18) holds with $V(x) := ||x||_2$. Heuristically, taking the random contribution (Γ_n) to approximate the delta function at the `0`-matrix results in a multidimensional `AR(1)` process that is under consideration. This implies that (heuristically) a counterpart of eq. (5.11) holds for a V-norm, i.e.:

$$||P^n(\mathscr{G}^{(0)},\cdot) - \pi(\cdot)||_V \leq R(m,T)V(\mathscr{G}^{(0)},m,T)(\rho(m,T)+\delta)^n.$$

For it, the form of the function V used on the right hand side of the inequality coincides with the choice of V for a V-norm ($V(x) := ||x||_2$) and thus: $||P^n(\mathscr{G}^{(0)},\cdot) - \pi(\cdot)||_V \leq R(m,T)||\mathscr{G}^{(0)}||_2(\rho(m,T)+\delta)^n$, for some function $R(m,T)$. In fact, it should be expected that the dependence on m comes only indirectly, through the dependence on $\mathscr{G}^{(0)}$, therefore it is reasonable to anticipate $R(m,T)$ to be only weakly dependent on m. Moreover, $||\mathscr{G}^{(0)}||_2 \leq \sqrt{m}\max\{|\mathscr{G}^{(0)[i]}|; i = 1,\dots,m\}$, therefore, at least for the convergence in V-norm, the statement of conjecture 5.2.1 seems particularly weak.

Corollary 5.2.3. Under conjecture 5.2.1, if $\mathbb{P}_b$ denotes the law of Brownian motion, the Markov chain defined by the blocked rejection sampler on a path space outputs an independent sample at an ϵ-tolerance level after $N(T,m,\epsilon,\mathscr{G}^{(0)})$ steps, with

$$\begin{aligned} N(T,m,\epsilon,\mathscr{G}^{(0)}) = {} & \Omega\big(m^2[\log(m)+\log(T)]\big) + \Omega\big(m^2[-\log(\epsilon)]\big) \\ & + \Omega\Big(m^2\Big|\log(f(\mathscr{G}^{(0)}))\Big|\Big), \quad \text{for } m,T \to \infty \text{ and } \epsilon \downarrow 0, \end{aligned}$$

with f being some appropriate, fixed function. If $\mathbb{P}_b$ denotes the law of the Ornstein-Uhlenbeck process, then the same holds true if $T = o(m)$.

Remark 5.2.1. In particular, it is possible that for the Ornstein-Uhlenbeck process a smaller $N(T,m,\epsilon,\mathscr{G}^{(0)})$ can be taken if m is not required to grow faster than T.

It now follows from corollary 5.2.3 that for a fixed $\mathscr{G}^{(0)}$, and a fixed, small enough ϵ, the computational cost of blocked rejection sampler on a path space can be bounded by:

$$\mathscr{R}_{blocking}(T,m) = \mathscr{O}(Tm^2(\log(m)+\log(T))\exp\{2cT/m\}), \quad \text{as } T,m \to \infty,$$

where the corresponding restriction of m growing at a faster rate than T is made whenever necessary.

It is now possible to make an informed decision about the choice of m. Clearly, as $T \to \infty$, the most worrying contribution comes from the term $\exp\{2cT/m\}$, thus to prevent an explosion of cost, m must be at least of the same order as T. On the other hand, m should not be any larger, due to contribution of the term $m^2\log(m)$. Consequently, setting $m = \Theta(T)$ under the assumption of $\mathbb{P}_b$ denoting the law of Brownian motion and $m = \Theta(T\log(T))$ under the assumption of the Ornstein-Uhlenbeck process results in a bound:

$$\mathscr{R}_{blocking}(T) = \mathscr{O}(T^3\log^3(T)), \quad \text{as } T \to \infty. \tag{5.12}$$

This shows that blocking can reduce the scaling of the computational cost of the algorithm from exponential in T to slightly in excess of cubic.

5.3 Non-centred parametrisation

The running application of this thesis is Bayesian inference for diffusion processes via data augmentation. For it, a non-centrally parametrised version of the conditioned diffusion sampler is needed (see sections 2.6 and 4.1.4 for an introduction to non-centred parametrisation). In this section I will show how a non-centrally parametrised rejection sampler on a path space with blocking can be formulated in two distinct ways. I follow the notation from section 2.6.

5.3.1 Multiple non-centred spaces

For the first approach, the centred probability spaces $(\Omega_i, \mathscr{F}_i^{(\theta)}, \mathbb{P}_i^{(\theta)})_{(\theta,i)}, (i = 1,\ldots,\Bbbk)$, $\theta \in \Theta$ are indexed not only with a parameter θ, but also with a knot index i. There are also $\Bbbk$ non-centred probability spaces—one for each set of knots—$(\Omega_i, \mathscr{F}_i, \mathbb{P}_i)$, $(i = 1,\ldots,\Bbbk)$, with their corresponding stochastic processes $W_i \sim \mathbb{Q}_i^*, (i = 1,\ldots,\Bbbk)$.

Finally, $\mathbb{k}$ functions $\Psi_{\theta,i} : \Omega_i^* \to \Omega$, $(i = 1,\ldots,\mathbb{k})$ are defined, for which the push-forward measures $(\mathbb{Q}_i^*)_{\#}(\Psi_{\theta,i})$ coincide with the laws $\mathbb{P}_i^{(\theta)}$, $(i = 1,\ldots,\mathbb{k})$. The goal is to define a Markov chain

$$\{(X^{(n)}, W_1^{(n)},\ldots, W_{\mathbb{k}}^{(n)}); n = 0,\ldots,\mathbb{k}N\},$$

whose invariant distribution is characterised by the following two properties:

- The invariant density of the marginal chain $\{X^{(\mathbb{k}n)}; n = 1,\ldots,N\}$ is equal to $\mathbb{P}^{(\theta)}$.
- The following identity holds $\Psi_{\theta,1}(W_1^{(n)}) = \cdots = \Psi_{\theta,\mathbb{k}}(W_{\mathbb{k}}^{(n)}) = X^{(n)}$ for each $n = 0,\ldots,\mathbb{k}N$.

Additionally, it is only permitted to sample from $\mathbb{Q}_i^*$, $i = 1,\ldots,\mathbb{k}$ (and any standard laws, such as `Unif([0,1])` or `Exp(1)`).

In the context of rejection sampling on a path space with blocking, the centred probability laws are given by the proposal diffusion law conditioned on the expanded conditioning set $\mathcal{Z}_{-i}$ (defined in eq. (5.1))

$$\mathbb{P}_i^{(\theta)} := \mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z}_{-i}).$$

Notice that the laws above, restricted to $\mathcal{E}_i$, $i = 1,\ldots,\mathbb{k}$, are used by the blocked rejection sampler on a path space in line 5 of algorithm 5.1 as proposals for $\mathbb{P}_\alpha^{(\theta)}|_{\mathcal{E}_i}(\cdot|\mathcal{Z}_{-i})$. For convenience, I refer to the elements of $\iota(\mathcal{Z}_{-i})$ as:

$$\iota(\mathcal{Z}_{-i}) := \{k_{ji}; j = 0,\ldots,\varkappa_i\}.$$

The non-centred probability spaces are defined through:

$$\begin{aligned}
&\Omega_i^* := \prod_{j=0}^{\varkappa_i - 1} \mathcal{C}([0,\Delta_j^{(i)}];\mathbb{R}^d), \quad \left(\Delta_j^{(i)} := k_{(j+1)i} - k_{ji}, \quad j = 0,\ldots,(\varkappa_i - 1)\right).\\
&\mathscr{F}_i^* := \mathscr{B}(\Omega_i^*), \qquad \mathbb{Q}_i^* = \bigotimes_{j=0}^{\varkappa_i - 1} \mathbb{W}^*[\Delta_j^{(i)}], \quad (i = 1,\ldots,\mathbb{k}),
\end{aligned} \tag{5.13}$$

where $\mathscr{B}$ denotes a Borel-σ-algebra and $\mathbb{W}^*[\Delta_j^{(i)}]$, $(i = 1,\ldots,\mathbb{k})$ denote the laws of independent, 0–0 Brownian bridges defined on the intervals $[0,\Delta_j^{(i)}]$. Correspondingly, the non-centred stochastic process W_i consists of $\varkappa_i$ independent, 0–0

Brownian bridges, which I denote with: $W_i := \{W_i^{[j]};\, j = 0,\dots,(\varkappa_i - 1)\}$. Finally, functions $\Psi_{\theta,i}$ are defined as:

$$\Psi_{\theta,i}(W) := \left\{ \sum_{j=0}^{\varkappa_i - 1} \Psi_{\theta,i}^{[j]}(W) \mathbb{1}_{(k_{ji}, k_{(j+1)i}]}(t);\, t \in [0,T] \right\}, \quad \text{where}$$

$$\Psi_{\theta,i}^{[j]}(W) := \left\{ W_{i,t-k_{ji}}^{[j]} + Y_{k_{ji}}(\theta)\left(1 - \frac{t-k_{ji}}{\Delta_j^{(i)}}\right) + Y_{k_{(j+1)i}}(\theta)\frac{t-k_{ji}}{\Delta_j^{(i)}};\, t \in \left[k_{ji}, k_{(j+1)i}\right] \right\}, \tag{5.14}$$

$(i = 1,\dots,\mathbb{k})$. To put it simply, for each $i = 1,\dots,\mathbb{k}$ the interval $[0,T]$ is chopped into smaller sub-intervals $[k_{ji}, k_{(j+1)i}]$, $j = 0,\dots,\varkappa_i - 1$ and on each one of those a 0–0 Brownian bridge is sampled. Then, function $\Psi_{\theta,i}(W)$ linearly translates sampled bridges, so that the end-points agree with the path values $\{Y_t(\theta);\, t \in \iota(\mathcal{Z}_{-i})\}$, akin to how it is done in fig. 4.1, but for an expanded set of observations $\mathcal{Z}_{-i}$ instead of the set of observations $\mathcal{Z}$. Algorithm 5.1 can now be modified as follows

Algorithm 5.2 Non-centrally parametrised blocked rejection sampler on a path space I

1: Draw $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z})$
2: Set $X \leftarrow \{\eta_\theta^{-1}(Y_t);\, t \in [0,T]\}$ ▷ In practice, this step can be omitted
3: **for** $n = 1,\dots,N$ **do**
4: **for** $i = 1,\dots,\mathbb{k}$ **do**
5: Draw $I \sim q_I(i,\cdot)$
6: `accepted` $\leftarrow$ `False`
7: **repeat** ▷ Equivalent to sampling $Y|_{\mathcal{E}_I} \sim \mathbb{P}_\alpha^{(\theta)}|_{\mathcal{E}_I}(\cdot|\mathcal{Z}_{-I})$
8: Draw $W_I|_{\mathcal{E}_I} \sim \mathbb{Q}_I^*|_{\mathcal{E}_I}$ ▷ See eq. (5.13)
9: Set $Y \leftarrow \Psi_{\theta,I}(W_I)$ ▷ See eq. (5.14)
10: Set `accepted` $\leftarrow$ `True`, with probability $\propto \frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y|\mathcal{Z}_{-I})$
11: **until** `accepted`
12: Set $X|_{\mathcal{E}_I} \leftarrow \{\eta_\theta^{-1}(Y_t);\, t \in \mathcal{E}_I\}$ ▷ In practice, this step can be omitted
13: Set $W_j \leftarrow \Psi_{\theta,j}^{-1}(Y)$, for $j \neq I$

A restriction to $\mathcal{E}_I$ in line 8 of algorithm 5.2 indicates that the sampling of 0–0 Brownian bridges needs to be done only on those intervals which belong to $\mathcal{E}_I$, in particular, W_I on $\mathcal{E}_I^C$ remains unaffected (see remark 5.3.1 for a minor exception to this convention). Notice that this does not impact the tractability of the law $\mathbb{Q}_I^*|_{\mathcal{E}_I}$. Let me also remark that in line 13 it is not necessary to compute all W_j with

$j \neq I$. This step can be temporarily put on hold, the algorithm can proceed to line 5 and a new index $\widetilde{I}$ can be sampled. Then, line 13 which was put on hold may be retrospectively executed by computing $W_{\widetilde{I}}$ only (notice that other W_j's are not used in any of the steps in lines 6–12 and once line 13 is reached again, all of them would have needed to be re-computed anyway).

Algorithm 5.2 is the most literal translation of the algorithm 5.1 to a non-centrally parametrised setting. This is also the approach taken by van der Meulen and Schauer (2018) in defining a non-centrally parametrised version of guided proposals with blocking. For the latter algorithm this seems to be the only reasonable solution, but for the rejection sampler on a path space with blocking there exists an alternative formulation that gives rise to slightly more elegant computer code.

5.3.2 Blocking directly on a non-centred space

For the second approach, the centred probability spaces $(\Omega_i, \mathscr{F}_i^{(\theta)}, \mathbb{P}_i^{(\theta)})_{(\theta,i)}$ are again indexed not only by $\theta \in \Theta$ but also by $i = 1, \ldots, \Bbbk$, although now, a single, non-centred probability space $(\Omega^*, \mathscr{F}^*, \mathbb{Q}^*)$, with a stochastic process W and a single function Ψ_θ are used. Additionally, certain random variables $\widetilde{\mathcal{Z}}_{-i}$, $i = 1, \ldots, \Bbbk$ need to be defined for which the conditioned pushforward measures satisfy

$$(\mathbb{Q}^*(\cdot|\widetilde{\mathcal{Z}}_{-i}))_\#(\Psi_\theta)(B) := \mathbb{Q}^*(\Psi_\theta^{-1}(B)|\widetilde{\mathcal{Z}}_{-i}) = \mathbb{P}_i^{(\theta)}(B), \quad \forall B \in \mathscr{F}_i^{(\theta)}. \tag{5.15}$$

In the context of blocked rejection sampling on a path space, this type of non-centred parametrisation can be succinctly described as blocking performed directly on the non-centred probability space $(\Omega^*, \mathscr{F}^*, \mathbb{Q}^*)$. Following this analogy, define the set of knots $\widetilde{\mathscr{K}}$ for the process W (instead of X), as well as related $\widetilde{\mathscr{K}_i}$, $\widetilde{\mathscr{K}_{-i}}$, $\widetilde{\mathcal{Z}}_{-i}$ and $\widetilde{\mathcal{E}}_i$ analogously to how it was done at beginning of section 5.1, by replacing X with W. Just as in section 5.3.1, define the centred probability spaces as $\mathbb{P}_i^{(\theta)} := \mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z}_{-i})$. This time however, let the non-centred probability space be given by:

$$(\Omega^*, \mathscr{F}^*, \mathbb{Q}^*) = \left(\mathcal{C}([0,T];\mathbb{R}^d), \mathscr{B}\left(\mathcal{C}([0,T];\mathbb{R}^d)\right), \mathbb{W}^*\right),$$

where $\mathbb{W}^*$ denotes the law induced by 0–0 Brownian bridges on $[0,T]$. Finally, define Ψ_θ via:

$$\Psi_\theta(W) := \left\{ W_t + Y_0(\theta)\left(1 - \frac{t}{T}\right) + Y_T(\theta)\frac{t}{T} \,;\, t \in [0,T] \right\}.$$

The bridges W—instead of being sampled in full—are updated in blocks, by sampling from $\mathbb{W}^*(\cdot|\widetilde{\mathcal{Z}}_i)$. Function Ψ_θ linearly translates paths W in such a way that the end-points agree with the end-points of Y. Notice that because the update $W \sim \mathbb{W}^*(\cdot|\widetilde{\mathcal{Z}}_i)$ holds the points W_t, $t \in \iota(\widetilde{\mathcal{Z}}_{-i})$ stationary, the points X_t, $t \in \iota(\mathcal{Z}_{-i})$ also remain unaffected during this procedure. It is thus clear that eq. (5.15) holds true. Algorithm 5.3 summarises these steps.

Algorithm 5.3 Non-centrally parametrised blocked rejection sampler on a path space II

Draw $Y \sim \mathbb{P}_0(\cdot|\mathcal{Z})$
Set $X \leftarrow \{\eta_\theta^{-1}(Y_t);\, t \in [0,T]\}$ ▷ In practice, this step can be omitted
for $n = 1,\dots,N$ **do**
  for $i = 1,\dots,\Bbbk$ **do**
    Draw $I \sim q_I(i,\cdot)$
    `accepted` $\leftarrow$ `False`
    repeat ▷ Equivalent to sampling $Y|_{\mathscr{E}_I} \sim \mathbb{P}_\alpha^{(\theta)}|_{\mathscr{E}_I}(\cdot|\mathcal{Z}_{-I})$
      Draw $W|_{\mathscr{E}_I} \sim \mathbb{Q}^*|_{\mathscr{E}_I}(\cdot|\widetilde{\mathcal{Z}}_I)$
      Set $Y \leftarrow \Psi_\theta(W)$
      Set `accepted` $\leftarrow$ `True`, with probability $\propto \frac{\mathrm{d}\mathbb{P}_\alpha}{\mathrm{d}\mathbb{P}_0}(Y|\mathcal{Z}_{-I})$
    until `accepted`
    Set $X|_{\mathscr{E}_I} \leftarrow \{\eta_\theta^{-1}(Y_t);\, t \in \mathscr{E}_I\}$ ▷ In practice, this step can be omitted

Remark 5.3.1. A step of parameter update for Bayesian inference for diffusion processes requires special care. Recall from eq. (4.11) that this step hinges upon the possibility to write down in closed form the joint density for the parameter θ, the data and a surrogate $\mathscr{S}$ for the entire imputed path. In particular, to achieve that, each path imputation step needs to update an entire unobserved path and not only $X|_{\mathscr{E}_i}$, $i = 1,\dots,\Bbbk$ (which would be sufficient if parameter θ were fixed). Consequently, when blocking is employed for Bayesian inference, all steps updating $W|_{\mathscr{E}_i}$, $i = 1,\dots,\Bbbk$ need to be substituted with the corresponding steps updating W on $[0,T]$. The modifications required for this change follow trivially.

5.4 Numerical examples

The arguments employed in section 5.2 to derive the asymptotic bound (5.12) break down when assumption A15 does not hold. Nonetheless, the inequality (5.12) ap-

pears to still hold empirically even for highly non-linear diffusions. To illustrate this (wider than anticipated) applicability, I will present detailed analysis of the computational cost of blocked rejection sampler on a path space for one of the notoriously difficult to sample, one-dimensional diffusion models, often used as a benchmark—bridges of the sine diffusion. Not only will I demonstrate that blocking renders feasible the problem of exact sampling of such a challenging process, I will also show that the adherence to the bound in eq. (5.12) is surprisingly faithful.

5.4.1 Sine example

Suppose that the target process solves the following stochastic differential equation:

$$\mathrm{d}X_t = (2 - 2\sin(8X_t))\,\mathrm{d}t + \frac{1}{2}\mathrm{d}W_t, \quad X_0 = 0, \quad t \in [0, T]. \tag{5.16}$$

As usual, I denote the law induced by it with $\mathbb{P}_b$. I consider six problems—increasing

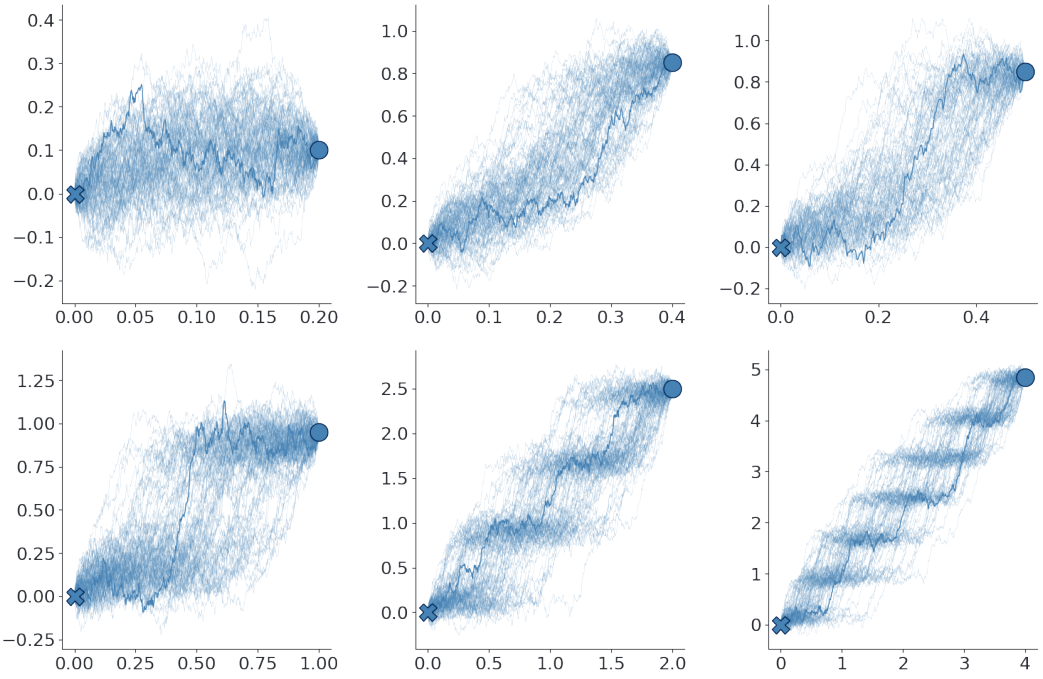


Figure 5.3: Typical paths of the sine diffusion in eq. (5.16), conditioned on an end-point: $X_{0.2} = 0.1$ (top-left), $X_{0.4} = 0.85$ (top-centre), $X_{0.5} = 0.85$ (top-right), $X_1 = 0.95$ (bottom-left), $X_2 = 2.5$ (bottom-centre), $X_4 = 4.85$ (bottom-right). Paths were simulated using rejection sampling on a path space with blocking. The number of knots used to produce the plots were chosen to be 0, 0, 0, 2, 12 and 40 respectively and for the latter three plots MCMC chain was run for 10^6 iterations—the 100 plotted paths come from thinning the latter half of these chains.

in difficulty—of simulating conditioned processes $\mathbb{P}_b(\cdot|\mathcal{Z})$, with $\mathcal{Z} = X_T$, when T is set to 0.2, 0.4, 0.5, 1, 2 or 4 and X_T is respectively equal to 0.1, 0.85, 0.85, 0.95, 2.5 or 4.85. Typical paths simulated under the respective laws are given in fig. 5.3. I chose the values of the end-points by simulating unconditioned paths of the SDE in eq. (5.16), plotting marginal distribution of the simulated paths at the given time-points and picking some points in the vicinity of modes.

For $T = 0.2$ paths resemble Brownian bridges, but as T increases, the non-linear dynamics becomes pronounced: the diffusion is effectively attracted to a ladder of values and it is repelled at the intermediate points, leading to a multimodal behaviour of the trajectories. As a result, it is virtually impossible to draw paths from the bottom row of fig. 5.3 using rejection sampling on a path space without blocking.

For each one of the six problems I drew 10^5 trajectories (in an MCMC setting) using blocked rejection sampling on a path space with `CBUS`, for various number of knots. For the first three problems I also employed regular rejection sampler, without blocking. I recorded the time required to sample a single path (which for a blocked sampler is counted as a single Gibbs sweep—lines 3–5 of the algorithm 5.1) and plotted it against the number of knots used in fig. 5.4 (knots were always placed at an equidistant grid). Notice that for $T = 0.2$, rejection sampling on a path space

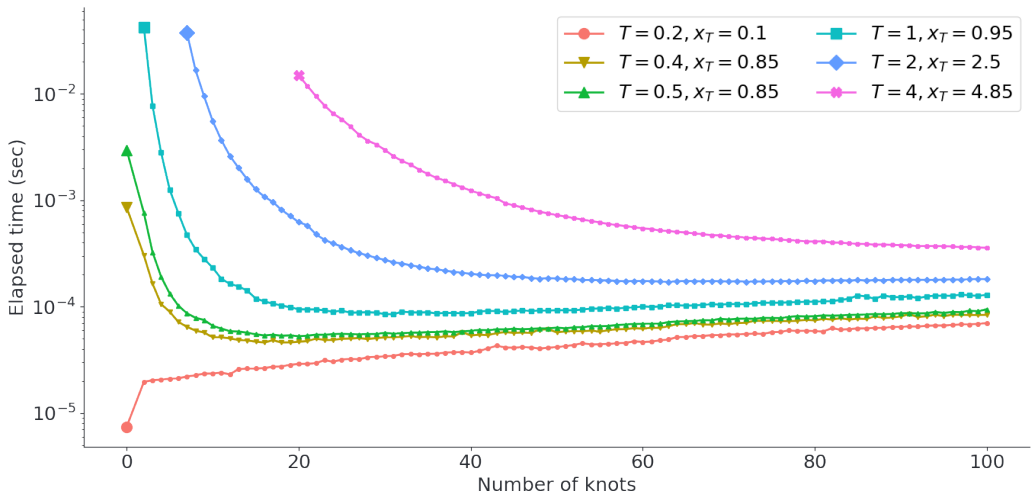


Figure 5.4: Time (in seconds) required to sample a single path of the sine diffusion from eq. (5.16) (for the blocked sampler it corresponds to a single Gibbs sweep—lines 3–5 of algorithm 5.1) as a function of the number of knots used (mind the log-scale on the y-axis).

clearly outperforms any blocking scheme, because paths under $\mathbb{P}_b(\cdot|\mathcal{Z})$ closely resemble Brownian bridges. Nonetheless, as T increases, this is no longer the case and blocking makes it possible to substantially reduce the cost of obtaining any single sample path. For instance, note the steep (exponential) reduction in cost for $T = 1$, $x_T = 0.95$. The same steep curve (and in fact much steeper ones) would have been drawn for the two most challenging setups: $x_2 = 0.95$ and $x_4 = 4.85$, had the experiments with smaller number of knots were run; however, due to time constraints executing such experiments would have taken prohibitively long execution times. This clearly illustrates cost reductions resulting from eq. (5.3).

Nevertheless, as argued in section 5.2, there is another cost that needs to be accounted for—correlation between sampled paths. The autocorrelation plots for $X^{(n)}_{t^*}$, when t^* coincides with the position of the most central knot (or in the case of an even number of knots: one of the two of the most central knots) given in fig. 5.5 illustrate this concept. Notice that for all examples as the number of knots

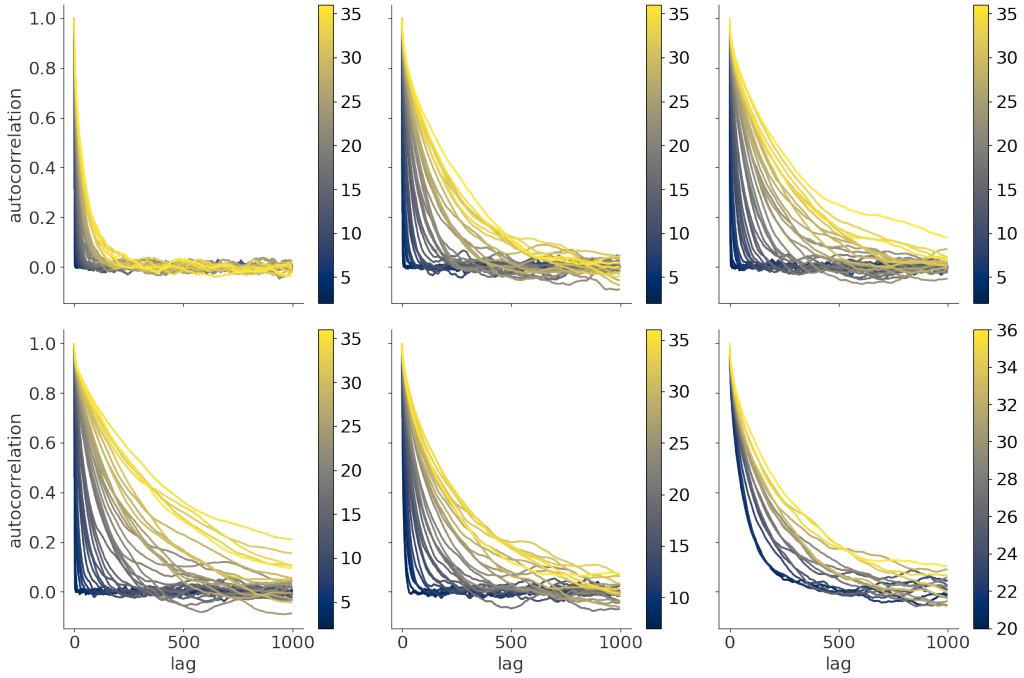


Figure 5.5: Autocorrelation plots for the sine example. Top-left: $T = 0.2$, $x_T = 0.1$; top-centre: $T = 0.4$, $x_T = 0.85$, top-right: $T = 0.5$, $x_T = 0.85$, bottom-left: $T = 1$, $x_T = 0.95$, bottom-centre: $T = 2$, $x_T = 2.5$, bottom-right: $T = 4$, $x_T = 4.85$. Each line corresponds to an experiment run with different number of knots used (number of knots corresponding to a colour used are given in the *colormaps* to the right of autocorrelation plots)

increases, the autocorrelation tapers off at a later lag. This means that even though each trajectory can be sampled quicker, a greater number of them needs to be drawn before an independent sample at an ϵ-tolerance level is obtained.

In order to quantify all of the involved costs it is thus better to examine fig. 5.6 below, relating time-adjusted effective sample size and the (half-) length of blocks.

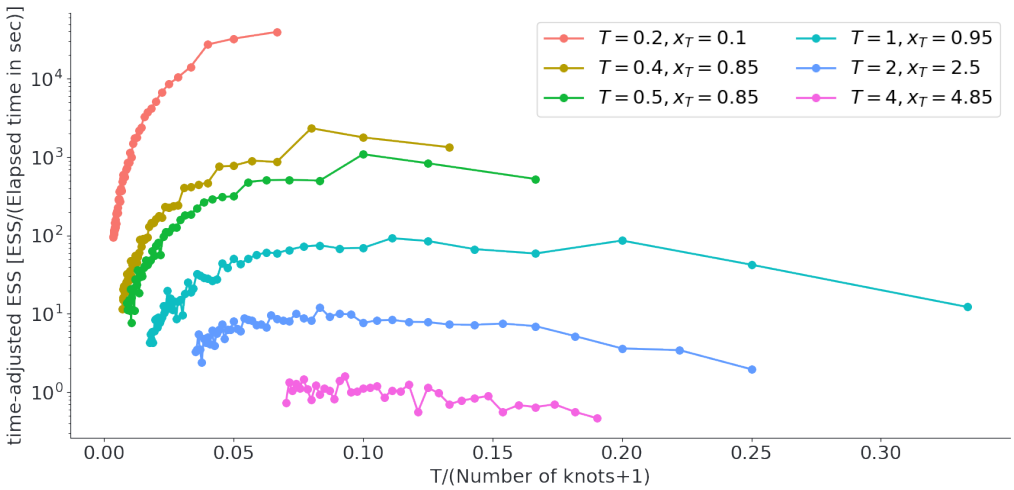


Figure 5.6: Time-adjusted effective sample size (taESS: (effective sample size)/(elapsed time in sec to sample entire chain)) vs half-length of blocks.

For computing the effective sample size (ESS) I used the function `effective_n` implemented in the package `PyMC3`. The time-adjusted effective sample size (taESS) is defined here as ESS divided by the amount of time taken to sample the entire chain $\{X^{(n)}, n = 1, \ldots, N\}$. Consequently, taESS is approximately equal to a number of independent samples that can be drawn in one second. Clearly, the larger the taESS, the more efficient the algorithm is.

It is important to note that for any sampling setting, there will always exist a point for which increasing the number of knots any further will only lead to a decrease in taESS (illustrated by the sharp dips of curves on the left side of the fig. 5.6). It is equally important to realise that for regimes with large enough T placing too few knots also leads to suboptimal strategies with lower than possible taESS. Even though the curves in fig. 5.6 clearly illustrate this concept by decreasing towards the right side of the graph, the degree of steepness that such declines can reach cannot be fully illustrated, simply because performing experiments needed for it would

take prohibitively long time. As a result, each problem admits a range of optimal values for the number of knots that yield the highest levels of taESS. Based on fig. 5.6 choosing the number of knots in such a way that the half-lengths of blocks are approximately equal to 0.1 seems to give uniformly superior performance.

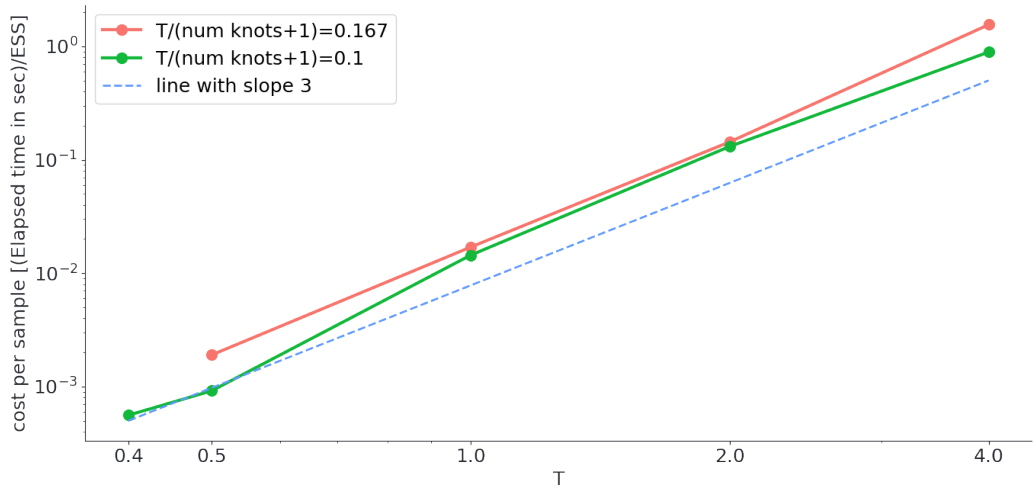


Figure 5.7: Computational cost as a function of time for the sine example. Inverse of time-adjusted effective sample size is used as a proxy for computational cost of the sampler.

As the last point, let me illustrate how the bound in eq. (5.12) can be verified in practice. First, notice that the inverse of taESS is approximately equal to the amount of time needed for obtaining a single sample that is independent at an ϵ-tolerance level. This is precisely how the computational cost of blocked rejection sampler on a path space has been defined in section 5.2. Equation (5.12) says that this cost as a function of T scales at most cubically (or to be precise is bounded by a term slightly in excess of cubic in T), so long as the number of knots is taken to be (almost) proportional to T. This means that for $T/(\text{number of knots}+1) \approx c$, for some constant $c \in \mathbb{R}_+$ taESS(T) should be at most a cubic function of T. If plotted on the log-log scale (where both axis are transformed to log-scales), eq. (5.12) is equivalent to taESS(T) being dominated by some line with slope 3. Figure 5.6 gives this precise plot, showing that in fact the bound is reached, because on log-log scale taESS(T) as a function of T appears to be a line with slope 3. This means that taESS$(T) \propto T^3$.

5.5 Discussion

In this chapter I quantified the asymptotic reduction in the computational cost of the exact rejection sampler on a path space offered by employment of the blocking technique. I showed that for the two simple choices of diffusion processes: Brownian motion and the Ornstein-Uhlenbeck process, this quantification can be made reasonably precise (though, due to conjecture 5.2.1 perhaps not to a satisfactorily rigorous extent). I showed that for these two examples the computational cost of a blocked rejection sampler on a path space scales at a rate that as a function of the duration of the bridges is at most (slightly in excess of) cubic. In the numerical section I considered a highly non-linear, scalar diffusion example and I showed that the theoretical upper bound that has been derived for Brownian motion and the Ornstein-Uhlenbeck process holds empirically also in this more elaborate setting.

The results offer certain heuristics for choosing the number of blocks. As pointed out in section 5.2.2, the number of knots ought to be chosen to be roughly proportional to the interval length, yielding constant *effective* length over which path-segments are ever imputed (say, if T is the interval length and the number of knots is m, then the effective length is $2T/(m+1)$, just like in the discussion preceding eq. (5.3)). The optimal effective length is not known a priori, but can be found experimentally for a single interval length and then used to find the optimal number of knots for any other interval length. In practice, one could depart from the assumption of an equidistant grid and investigate the problem of the optimal placement of knots.

As argued in chapter 4, simulation of conditioned diffusions is often just a component of more composite algorithms. In such scenarios it is the impact that blocking has on the cost of the overall procedure that is of interest to a practitioner. For instance, in the setting of Bayesian inference for diffusion processes via data augmentation, path updates constitute just one half of all the algorithmic steps. Another half comprise of parameter updates. Additionally, it is the mixing time of the parameter chain $\theta^{(n)}$ and not the path $X^{(n)}$ that is of interest. It would be of great practical interest to extend the results of this chapter to the characterisation of the computational cost of the parameter chain.

Proofs

Proof of lemma 5.2.1. $\mathcal{G}$ under assumption A15 is a special case of the Gaussian process considered by Roberts and Sahu (1997). Therefore, the statement of Roberts and Sahu (1997, Theorem 1) applies and the convergence rate is given by $\rho_{\texttt{spec}}(B)$. Due to tridiagonal structure of the precision matrix Λ (which follows from the Markov property of the process $\mathcal{G}$, see also the proof of theorem 5.2.1), Roberts and Sahu (1997, Corollary 3) applies, meaning that the $\mathcal{L}^2$ convergence rate of `CBUS` is equal to that of `LBUS`. By the same token, Roberts and Sahu (1997, Theorem 5) applies as well, yielding $\rho_{\texttt{spec}}(B) = \lambda^2_{\texttt{max}}(A)$ under `CBUS` and `LBUS`. Finally, the $\mathcal{L}^2$ convergence rate of `RBUS` follows from Roberts and Sahu (1997, Theorem 2). □

Proof of theorem 5.2.1. The precision matrix Λ of any random vector $\mathcal{G}$ with non-degenerate covariance matrix can be related to the matrix of partial correlations via (Lauritzen, 1996, p. 130):

$$\mathbb{C}orr(\mathcal{G}^{[i]}, \mathcal{G}^{[j]} | \mathcal{G} \setminus \{\mathcal{G}^{[i]}, \mathcal{G}^{[j]}\}) = -\frac{\Lambda^{[i,j]}}{\sqrt{\Lambda^{[i,i]}\Lambda^{[j,j]}}}.$$

By the definition of $\mathcal{G}$ in eq. (5.4), it is easy to see that $\mathbb{C}orr(\mathcal{G}^{[i]}, \mathcal{G}^{[j]} | \mathcal{G} \setminus \{\mathcal{G}^{[i]}, \mathcal{G}^{[j]}\}) = 0$ whenever $|i-j| > 1$; that by symmetry $\Lambda^{[i,i+1]} = \Lambda^{[i+1,i]}$, $(i = 1, \ldots, m)$; and that $\Lambda^{[i,i]} = \Lambda^{[j,j]}$, $(i, j = 1, \ldots, m)$, because $\mathbb{V}ar(\mathcal{G}^{[i]} | \mathcal{G} \setminus \mathcal{G}^{[i]}) = (\Lambda^{[i,i]})^{-1}$, $(i = 1, \ldots, m)$ (Roberts and Sahu, 1997, p.296). In addition, under assumption A15, the covariance matrix depends only on time and not on the state variable, thus under the assumption of the equidistant grid $\mathbb{C}orr(\mathcal{G}^{[i]}, \mathcal{G}^{[i+1]} | \mathcal{G} \setminus \{\mathcal{G}^{[i]}, \mathcal{G}^{[i+1]}\}) =: c(\Delta)$, $(i = 1, \ldots, m-1)$. Consequently Λ is a Toeplitz matrix whose non-zero entries are related via $\Lambda^{[i,i+1]} = \Lambda^{[i+1,i]} = -\Lambda^{[i,i]} c(\Delta)$, $(i = 1, \ldots, m)$. The form of the matrix A follows:

$$A = \begin{pmatrix} 0 & c(\Delta) & 0 & \ldots & 0 \\ c(\Delta) & 0 & c(\Delta) & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & c(\Delta) & 0 & c(\Delta) \\ 0 & \ldots & 0 & c(\Delta) & 0 \end{pmatrix}. \tag{5.17}$$

By Smith (1985) (or alternatively Kulkarni et al. (1999)) the eigenvalues of matrix A are given by

$$-2c(\Delta)\cos\left(\frac{\pi l}{m+1}\right), \qquad l = 1, \ldots, m.$$

Depending on the sign of $c(\Delta)$ the maximal eigenvalue of A is therefore given by:

$$\lambda_{\texttt{max}}(A) = \begin{cases} -2c(\Delta)\cos\left(\frac{\pi m}{m+1}\right) = 2c(\Delta)\cos\left(\frac{\pi}{m+1}\right), & \text{if } c(\Delta) > 0, \\ -2c(\Delta)\cos\left(\frac{\pi}{m+1}\right) & \text{if } c(\Delta) < 0, \end{cases}$$

and the result concerning $\lambda_{\texttt{max}}(A)$ follows. The statement about convergence rates follows as well by substituting the expression for $\lambda_{\texttt{max}}(A)$ into lemma 5.2.1. □

Proof of corollary 5.2.1. By theorem 5.2.1 only $c(\Delta)$ needs to be computed. This can be done using standard properties of Brownian motion and Brownian bridges:

$$c(\Delta) = \frac{\mathbb{C}ov\left(X_\Delta, X_{2\Delta} | X_0, X_{3\Delta}\right)}{\sqrt{\mathbb{V}ar(X_\Delta | X_0, X_{3\Delta})\mathbb{V}ar(X_{2\Delta} | X_0, X_{3\Delta})}} = \frac{\frac{1}{3}\Delta}{\sqrt{\left(\frac{2}{3}\Delta\right)^2}} = \frac{1}{2}. \tag{5.18}$$

The asymptotic convergence (as $m \to \infty$) of $\rho_{\texttt{CBUS}}$ and $\rho_{\texttt{LBUS}}$ in eq. (5.7) follows immediately from the Taylor expansion of $\cos^2(x)$ around 0. For asymptotic convergence of $\rho_{\texttt{RBUS}}$, notice that by Taylor expansions of $\cos(x)$ around 0, $\log(1-x)$ around 0 and $\exp(x)$ around 0 respectively:

$$\begin{aligned} \rho_{\texttt{RBUS}} &= \exp\left\{ m \log\left[m^{-1}\left\{ m - 1 + \cos\left(\frac{\pi}{m+1}\right)\right\}\right]\right\} \\ &= \exp\left\{ m \log\left[1 - \frac{1}{m}\left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-5})\right\}\right]\right\} \\ &= \exp\left\{ -\left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4})\right\} \\ &= 1 - \left(\frac{\pi}{m+1}\right)^2 + \mathcal{O}(m^{-4}). \end{aligned}$$

□

Proof of corollary 5.2.2. $c(\Delta)$ in eq. (5.10) follows by direct substitution of relevant terms from eq. (5.9) into the definition of the partial correlation (used for instance in eq. (5.18)). To derive the asymptotic convergence rates (as $m \to \infty$), I use Taylor expansions to compute the following identity

$$\begin{aligned} e^{-a\Delta}\sinh(b\Delta)\sinh(d\Delta) = &\left(1 - a\Delta + \frac{1}{2}(a\Delta)^2 - \frac{1}{6}(a\Delta)^3 + \mathcal{O}(\Delta^4)\right) \\ &\cdot \left(b\Delta + \frac{1}{6}(b\Delta)^3 + \mathcal{O}(\Delta^5)\right)\left(d\Delta + \frac{1}{6}(d\Delta)^3 + \mathcal{O}(\Delta^5)\right) \end{aligned}$$

$$
\begin{aligned}
&= \left(1 - a\Delta + \frac{1}{2}(a\Delta)^2 - \frac{1}{6}(a\Delta)^3 + \mathcal{O}(\Delta^4)\right)\left(bd\Delta^2 + \frac{1}{6}(b^3d + bd^3)\Delta^4 + \mathcal{O}(\Delta^6))\right) \\
&= bd\Delta^2 - abd\Delta^3 + \left(\frac{1}{6}\left(b^3d + bd^3\right) + \frac{1}{2}a^2bd\right)\Delta^4 \\
&\qquad - \frac{1}{6}\left(ab^3d + abd^3 + a^3bd\right)\Delta^5 + \mathcal{O}(\Delta^6),
\end{aligned}
$$

from which it follows that

$$
\begin{aligned}
c(\Delta) = &\Big[\left(3\Delta^2 - 6\Delta^3 + 11\Delta^4 - 14\Delta^5 + \mathcal{O}(\Delta^6)\right) \\
&\qquad - \left(2\Delta^2 - 6\Delta^3 + \frac{32}{3}\Delta^4 - 14\Delta^5 + \mathcal{O}(\Delta^6)\right)\Big] \\
&\cdot \left[\left(3\Delta^2 - 3\Delta^3 + \frac{13}{2}\Delta^4 - \frac{11}{2}\Delta^5 + \mathcal{O}(\Delta^6)\right)\right. \\
&\qquad \left. - \left(\Delta^2 - 3\Delta^3 + \frac{29}{6}\Delta^4 - \frac{11}{2}\Delta^5 + \mathcal{O}(\Delta^6)\right)\right]^{-1/2} \\
&\cdot \Big[\left(6\Delta^2 - 12\Delta^3 + 25\Delta^4 - 34\Delta^5 + \mathcal{O}(\Delta^6)\right) \\
&\qquad - \left(4\Delta^2 - 12\Delta^2 + \frac{70}{3}\Delta^4 - 34\Delta^5 + \mathcal{O}(\Delta^6)\right)\Big]^{-1/2} \\
&= \frac{\Delta^2 + \frac{1}{3}\Delta^4 + \mathcal{O}(\Delta^6)}{2\Delta^2 + \frac{5}{3}\Delta^4 + \mathcal{O}(\Delta^6)} = \frac{1}{2} - \frac{3\Delta^2 + \mathcal{O}(\Delta^4)}{12 + 10\Delta^2 + \mathcal{O}(\Delta^4)}.
\end{aligned}
$$

This yields the asymptotic convergence rates $\rho_{\texttt{CBUS}}$ and $\rho_{\texttt{LBUS}}$:

$$
\begin{aligned}
4c^2(\Delta)\cos^2\left(\frac{\pi}{m+1}\right) &= 4\left(\frac{1}{2} - \frac{3\Delta^2 + \mathcal{O}(\Delta^4)}{12 + 10\Delta^2 + \mathcal{O}(\Delta^4)}\right)^2\left(1 - \left(\frac{\pi\Delta}{T}\right)^2 + \mathcal{O}(\Delta^4)\right) \\
&= 1 - \Delta^2\left[\frac{\pi^2}{T^2} + \frac{6}{6 + 5\Delta^2 + \mathcal{O}(\Delta^4)}\right] + \mathcal{O}(\Delta^4).
\end{aligned}
$$

Asymptotic convergence rate of $\rho_{\texttt{RBUS}}$ follows similarly:

$$
\begin{aligned}
\rho_{\texttt{RBUS}} &= \exp\left\{m\log\left[m^{-1}\left\{m - 1 + \left(1 - \frac{3\Delta^2 + \mathcal{O}(\Delta^4)}{6 + 5\Delta^2 + \mathcal{O}(\Delta^4)}\right)\left(1 - \frac{1}{2}\left(\frac{\pi\Delta}{T}\right)^2 + \mathcal{O}(\Delta^4)\right)\right\}\right]\right\} \\
&= \exp\left\{m\log\left[1 - \frac{1}{m}\Delta^2\left[\frac{\pi^2}{2T^2} + \frac{3}{6 + 5\Delta^2 + \mathcal{O}(\Delta^4)}\right] + \frac{1}{m}\mathcal{O}(\Delta^4)\right\}\right]\right\} \\
&= \exp\left\{-\Delta^2\left[\frac{\pi^2}{2T^2} + \frac{3}{6 + 5\Delta^2 + \mathcal{O}(\Delta^4)}\right] + \mathcal{O}(\Delta^4)\right\} \\
&= 1 - \Delta^2\left[\frac{\pi^2}{2T^2} + \frac{3}{6 + 5\Delta^2 + \mathcal{O}(\Delta^4)}\right] + \mathcal{O}(\Delta^4).
\end{aligned}
$$

□

Proof of theorem 5.2.2. First, define:

$$L^2(\pi) := \left\{ \mu : \pi\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right)^2 \right] < \infty \right\},$$

and notice that for any $\mu \in L^2(\pi)$:

$$||\mu||^2_{L^2(\pi)} := \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) \right)^2 \pi(\mathrm{d}x) = \left\| \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|_{\mathcal{L}^2(\pi)}.$$

By Papaspiliopoulos et al. (prep, Proposition 14.9.2):

$$||P^n f(\mathscr{G}^{(0)}) - \pi(f)||_{\mathcal{L}^2(\pi)} = ||\mu(P^*)^n - \pi||_{L^2(\pi)}, \tag{5.19}$$

where μ is defined through the Radon-Nikodým derivative $f := \mathrm{d}\mu/\mathrm{d}\pi$ a.s., P^* denotes the adjoint transition operator and $\mu P := \int \mu(\mathrm{d}x) P(x, \cdot)$. Consequently the statement about $\mathcal{L}^2$ convergence rate found in definition 5.2.2 can be written equivalently as: for all $\delta > 0$ and for all measures $\mu \in L^2(\pi)$:

$$||\mu(P^*)^n - \pi||_{L^2(\pi)} \leq V_\mu(m, T)(\rho(m, T) + \delta)^n. \tag{5.20}$$

By assumption, the chain is reversible, so $P^* = P$. Therefore, if the $\mathcal{L}^2$ convergence rate of the chain equals $\rho(m, T)$, by the Cauchy-Schwarz inequality:

$$||\mu(P)^n - \pi||_{TV} \leq ||\mu(P^*)^n - \pi||_{L^2(\pi)} \leq V_\mu(m, T)(\rho(m, T) + \delta)^n.$$

Papaspiliopoulos et al. (prep, proof of Proposition 14.9.2) then implies:

$$||P^n(\mathscr{G}^{(0)}, \cdot) - \pi(\cdot)||_{TV} \leq V^\circ(\mathscr{G}^{(0)}, m, T)(\rho(m, T) + \delta)^n, \tag{5.21}$$

for all $\delta > 0$, some function V° and π-almost all $\mathscr{G}^{(0)}$. It is left to show that eq. (5.21) remains valid for all starting $\mathscr{G}^{(0)}$. This follows easily from the smoothness of the Gaussian kernels. In particular, since P is Gaussian, P^n is the n-convolution of Gaussian kernels, and thus Gaussian and smooth itself. Fix δ, n and $\mathscr{G}^{(0)}$. By smoothness, there exists $\widetilde{\delta} := \widetilde{\delta}(m, T) > 0$ s.t. for all $\widetilde{\mathscr{G}}^{(0)}$ with $d(\widetilde{\mathscr{G}}^{(0)}, \mathscr{G}^{(0)}) < \widetilde{\delta}$:

$$||P^n(\widetilde{\mathscr{G}}^{(0)}, \cdot) - P^n(\mathscr{G}^{(0)}, \cdot)||_{TV} \leq V^\circ(\mathscr{G}^{(0)}, m, T)(\rho(m, T) + \delta)^n.$$

Because a $\widetilde{\delta}$-ball around $\mathscr{G}^{(0)}$: $B_{\widetilde{\delta}}(\mathscr{G}^{(0)})$, has positive measure under π, there exists a starting vector $\widehat{\mathscr{G}}^{(0)} \in B_{\widetilde{\delta}}(\mathscr{G}^{(0)})$ for which (5.21) holds and thus:

$$||P^n(\mathscr{G}^{(0)}, \cdot) - \pi(\cdot)||_{TV} \leq ||P^n(\mathscr{G}^{(0)}, \cdot) - P^n(\widehat{\mathscr{G}}^{(0)}, \cdot)||_{TV} +$$
$$||P^n(\widehat{\mathscr{G}}^{(0)}, \cdot) - \pi(\cdot)||_{TV} \leq 2V^\circ(\mathscr{G}^{(0)}, m, T)(\rho(m, T) + \delta)^n,$$

and hence the statement of the theorem holds with $V := 2V^\circ$. □

Proof of corollary 5.2.3. Fix T, m and ϵ. Then, by theorem 5.2.2, a Markov chain defined by the blocked rejection sampler on a path space outputs an independent sample at an ϵ-tolerance level if N (the total number of steps of this Markov chain) is such that

$$V(\mathscr{G}^{(0)}, m, T)(\rho(m,T)+\delta)^N < \epsilon.$$

This is equivalent to:

$$N > -\frac{\log\big(V(\mathscr{G}^{(0)}, m, T)\big) - \log(\epsilon)}{\log(\rho(m,T)+\delta)}. \tag{5.22}$$

Set $\delta = 1/(m+1)^2$. By corollaries 5.2.1 and 5.2.2 and the Taylor expansion of $\log(1-x)$ around 0, eq. (5.22) is implied by:

$$N > \left(\frac{m+1}{a_1}\right)^2 \big(\log\big(V(\mathscr{G}^{(0)}, m, T)\big) - \log(\epsilon)\big),$$

with an appropriate constant a_1. Under conjecture 5.2.1, the above is implied by:

$$N > \left(\frac{m+1}{a_1}\right)^2 \left\{a_2\big[\log(1+m)+\log(1+T)+\log(f(\mathscr{G}^{(0)}))\big] - \log(\epsilon)\right\},$$

for some appropriate constant a_2. Consequently, for a fixed ϵ, the number of steps that need to be taken by the Markov chain to output an independent sample at an ϵ-tolerance level is:

$$N(T, m, \epsilon, \mathscr{G}^{(0)}) = \Omega\big(m^2[\log(m)+\log(T)]\big) + \Omega\big(m^2[-\log(\epsilon)]\big) + \Omega\Big(m^2\Big|\log(f(\mathscr{G}^{(0)}))\Big|\Big). \qquad \square$$

6 *Automation of Guided Proposals*

This chapter is based on joint work with Moritz Schauer and Frank van der Meulen.

Arguably, the three of the most important facets of any simulation algorithm are: its applicability (and closely related—robustness), its efficiency and its simplicity. Indeed, take the most widely used algorithm for sampling unconditioned diffusions—the Euler-Maruyama scheme. It is the three features above that ultimately led to its popularity: weak assumptions on the form of the drift and volatility coefficients guarantee broad applicability of this method to a wide range of diffusion models; the only computational steps comprise of matrix additions and multiplications, both of which are highly efficient and make it possible for the algorithm to be exceptionally fast; finally, computer code implementing the routine can be written even with a very limited knowledge of computer programming in a matter of minutes and within a few lines of code.

The problem of sampling conditioned diffusions is substantially more difficult than of simulating unconditioned ones and to the best of my knowledge there exist no algorithm suitable for this setting that would excel in all three domains above as emphatically as the Euler-Maruyama scheme does. With that being said, in this chapter I will show that guided proposals of Schauer et al. (2017) admit a certain formulation for which all three factors above—applicability, efficiency and simplicity—can be listed as its strengths.

As discussed in section 3.3.4, the applicability of guided proposals comes as a derivative of the weak assumptions imposed on the underlying diffusion. In particular, the restrictive assumptions A6 or A7 imposed by some of the competing methods do not apply, and thus the algorithm is particularly well-suited for multidimensional processes. An additional degree of flexibility awarded by the freedom over the choice of the auxiliary process means that guided proposals scale particularly well to sparse observations and high-dimensional settings. Additionally, even some hypoelliptic diffusions (when A3 does not hold but A4 does) can be tackled

with this method. As shown by van der Meulen and Schauer (2017b) and as I show in a different setting in chapter 7, various forms of conditioning can also be used. The reasons above imply that the range of diffusion-simulation problems that can be solved with guided proposals is strikingly broad.

The other two properties: efficiency and simplicity may not be as apparent. In fact, the way in which Schauer et al. (2017) compute the terms $\widetilde{r}$ and $\widetilde{H}$—which are defined in proposition 3.3.1 and are required for simulation of proposal paths—as well as $\widetilde{h}(0, x_0)$—which is defined in section 3.3.2 and is needed for evaluations of the likelihood functions—all involve repeated, expensive evaluations of matrix exponentials. Therefore, at least at the stage of the formulation of Schauer et al. (2017), efficiency of guided proposals could be put into question. To a certain extent, this has changed with a revamped formulation of van der Meulen and Schauer (2017b) in which $\widetilde{r}$ and $\widetilde{H}$ were shown to be related to the solutions of a system of certain backward ordinary differential equations. As a result, $\widetilde{r}$ and $\widetilde{H}$ could be computed by relying solely on the matrix addition, multiplication and inverse operations. This was an important improvement over the original approach of Schauer et al. (2017), but the formulation of van der Meulen and Schauer (2017b) is burdened with certain shortcomings. First, under many sampling regimes of interest (say, when the target diffusion is hypoelliptic), the dimension of the solution to the system of ordinary differential equations proposed by van der Meulen and Schauer (2017b) increase with the dimension of the dataset[1]. Second, the problem of accelerating computations of $\widetilde{h}$ has not been addressed and thus in the setting of statistical inference—where repeated evaluations of $\widetilde{h}$ are necessary—the only (fully general) way of obtaining $\widetilde{h}$ relies on performing multiple, expensive matrix multiplications.

In this chapter I will define a new set of backward ordinary differential equations which exist for all settings for which guided proposals are defined and which have a property that the dimension of their solution is constant in the size of the dataset. All of the $\widetilde{r}$, $\widetilde{H}$ and $\widetilde{h}$ can be derived from the solutions to these ODEs by means of matrix additions and multiplications and scalar operations. This means that simulating proposal paths consists of the following two steps:

[1]This can be managed to a certain extend by employing blocking—see van der Meulen and Schauer (2018) for details

- Solving a system of backward ODEs, so as to compute $\widetilde{r}$, $\widetilde{H}$ and $\widetilde{h}$ (notice that the computational cost of this step has the same order of magnitude as a forward simulation using the Euler-Maruyama scheme—in fact, in principle, the deterministic Euler scheme may be used to solve these ODEs[2]).

- Simulating an unconditioned proposal path defined by the SDE in eq. (3.19), using, say, the Euler-Maruyama scheme.

Consequently, simulating any proposal path is very efficient (it has the same order of cost as an unconstrained, forward simulation). This is not yet to say that the overall algorithm is efficient—for that it still needs to be shown that it is enough to simulate only a handful of such proposal paths. However, this follows immediately after recalling that guided proposals award the possibility of crafting better proposal laws by picking more fitting auxiliary processes. Because of the generic way in which $\widetilde{r}$, $\widetilde{H}$ and $\widetilde{h}$ are computed there is no price to pay for even the most elaborate choices, so long as they belong to the family of linear diffusions defined in eq. (3.20). Therefore "good" proposal laws are free from any computational overhead and consequently, an MCMC algorithm on a path space correcting for the discrepancies between the target ($\mathbb{P}_b(\cdot|\mathcal{Z})$) and the proposal ($\mathbb{P}_{b^\circ}$) laws need not take many steps before an independent sample at an ϵ-tolerance level is outputted (using definition 5.2.1). In fact, as shown in van der Meulen and Schauer (2017b), even in highly non-linear settings the auxiliary law may be chosen carefully enough so that the proposal law faithfully approximates the target. As a result, the overall computational cost of guided proposals (as compared to competing algorithms) is low.

Finally, there are two arguments to be made in favour of the claim about simplicity of guided proposals. First, once a generic version of the algorithm is implemented, any concrete implementation for a particular diffusion model boils down to two simple steps

- defining two functions—one evaluating the drift and another evaluating the volatility coefficient of the target process

[2]In practice, due to reasons explained below it is much better to use some higher-order Runge-Kutta schemes. I emphasise that their cost is still of the same order as the Euler scheme.

- specifying two other functions defined analogously for the auxiliary process

All of the remaining computations can be handled by a well-designed, generic implementation. Such implementation is currently available in `Julia` programming language upon installation of two packages: `Bridge.jl` and `BridgeSDEInference.jl`.[3] There is also a second argument speaking in support of the simplicity of guided proposals and that is that (a bare-bone version of) the algorithm is conceptually straightforward to implement and can be written by anyone familiar with computer programming without having to know the sophisticated mathematical machinery that justifies validity of all the steps. The three main ingredients are: solving a system of ordinary differential equations (from which $\widetilde{r}$, $\widetilde{H}$ and $\widetilde{h}$ can be derived) with standard techniques for ODEs; forward simulating proposal paths via the Euler-Maruyama scheme; and employing the Metropolis-Hastings correction by computing log-likelihoods via left-Riemann sum approximations to the integrals.

On top of presenting the new approach for deriving $\widetilde{r}$, $\widetilde{H}$ and $\widetilde{h}$, in this chapter I will also give a detailed explanation on how to generically handle uncertainty about the value of the starting point. One solution has been given in van der Meulen and Schauer (2017b); however, it suffers from the same computational issues as derivations of $\widetilde{r}$ and $\widetilde{H}$ do. The novelty of the approach presented in this chapter lies in the fact that no such increase in cost ever needs to be present.

6.1 Backward ordinary differential equations

In this section I provide a new set of backward ordinary differential equations that can be used to compute $\widetilde{r}$, $\widetilde{H}$ and $\widetilde{h}$. For completeness, let me start with repeating the relevant equations from the previous chapters.

The auxiliary process $\widetilde{X}$ is defined as the solution to the following stochastic differential equation

$$\mathrm{d}\widetilde{X}_t = \widetilde{b}(t,\widetilde{X}_t)\,\mathrm{d}t + \widetilde{\sigma}_t\,\mathrm{d}W_t, \quad \widetilde{X}_0 = x_0, \quad t \in [0,T],$$
$$\text{where } \widetilde{b}(t,x) := \widetilde{B}_t x + \widetilde{\beta}_t,$$

[3] I have primarily contributed to the `BridgeSDEInference.jl` package (`10.5281/zenodo.3446184`). Please see the github pages for a full and up-to-date list of all the contributors.

where $\widetilde{\beta}$, $\widetilde{B}_t$ and $\widetilde{\sigma}_t$ may all depend on time and $\widetilde{\Gamma}_t := \widetilde{\sigma}_t \widetilde{\sigma}_t^T$. Denote by $\widetilde{\mathbb{P}}$ the law induced by the SDE above. Additionally, define

$$\widetilde{h}(t,x) := \mathrm{d}\widetilde{\mathbb{P}}(\mathcal{Z}|\widetilde{X}_t = x),$$

where either $\mathcal{Z} := \widetilde{X}_T$ (in which case $\mathrm{d}\widetilde{\mathbb{P}}(\mathcal{Z}|\widetilde{X}_t = x) = \widetilde{p}(t,T,x,X_T)$ denotes a transition density of the auxiliary process $\widetilde{X}$) or $\mathcal{Z} := \{L_i \widetilde{X}_{t_i} + \xi_i,\, i = 1,\dots,K\}$, with ξ_i independent Gaussian random variables with mean 0 and covariance matrix Σ_i (in which case:

$$\mathrm{d}\widetilde{\mathbb{P}}(\mathcal{Z}|\widetilde{X}_t = x) = \int_{\mathscr{X}} \cdots \int_{\mathscr{X}} \widetilde{p}(t,t_1,x,x_1)\mathcal{N}_{\Sigma_K}(v_K - L_K x_K) \cdot \prod_{i=1}^{K-1} \Big[\widetilde{p}(t_i,t_{i+1},x_i,x_{i+1})\mathcal{N}_{\Sigma_i}(v_i - L_i x_i)\Big]\, \mathrm{d}x_1 \cdots \mathrm{d}x_K,$$

with $\mathcal{N}_{\Sigma_i}$ denoting the pdf of a multivariate Gaussian random variable, centred at zero and with a covariance matrix Σ_i ($i = 1,\dots,K$)).

In fact, as was shown in van der Meulen and Schauer (2017b, Remark 2.11) the results derived for the former conditioning are a special case of the latter one with the choice $K = 1$, $L_1 = I$ and $\Sigma_1 \downarrow 0$. Therefore, without loss of generality, I henceforth assume that

$$\mathcal{Z} := \{L_i \widetilde{X}_{t_i} + \xi_i,\, i = 1,\dots,K\}. \tag{6.1}$$

Now, define:

$$\widetilde{r}(t,x) := \nabla_x \log \widetilde{h}(t,x), \quad \text{and} \quad \widetilde{H}(t) := -D^2 \log \widetilde{h}(t,x),$$

where $D^2_{ij} f(t,x) := \partial^2 f(t,x)/(\partial x_i \partial x_j)$ and where independence of $\widetilde{H}(t)$ from the state variable can be verified by direct computations.

> The aim is to evaluate $\widetilde{h}(0,x_0)$ and compute $\widetilde{r}(t,x)$ and $\widetilde{H}(t)$ on any pre-specified time grid $0 = t_0 < \cdots < t_M < T$ and for any set of corresponding state variables $x_0,\dots,x_M$. The three objects are the only non-trivial unknowns needed for an implementation of guided proposals.

As observed by Schauer et al. (2017, §5) (and justified for instance in Jacobsen (1991)), $\widetilde{X}$ is a Gaussian process with mean $\mu_t(s,x) := \mathbb{E}[\widetilde{X}_t|\widetilde{X}_s = x]$ and covariance $K_t(s) := \mathbb{C}ov(\widetilde{X}_s, \widetilde{X}_t)$:

$$\mu_t(s,x) = \Phi(t,s)x + \int_s^t \Phi(s,\tau)\widetilde{\beta}_\tau \,\mathrm{d}\tau, \qquad K_t(s) := \int_s^t \Phi(t,\tau)\widetilde{\Gamma}_\tau \Phi(t,\tau)^T \,\mathrm{d}\tau, \tag{6.2}$$

where $\Phi(t,s) := \Phi(t)\Phi(s)^{-1}$ and $\Phi(t) \in \mathbb{R}^{d\times d}$, $t \in [0,T]$ is a fundamental matrix solution to:

$$\Phi(t) = I + \int_0^t \widetilde{B}_\tau \Phi(\tau)\,\mathrm{d}\tau.$$

For notational clarity it is easier to illustrate the results of van der Meulen and Schauer (2017b) and derive the novel backward ordinary differential equations under the assumption that the conditioning in eq. (6.1) comprises of only two observations:

$$\mathcal{Z} := \left\{ L_S\widetilde{X}_S + \xi_S \text{ and } L_T\widetilde{X}_T + \xi_T \right\}, \tag{6.3}$$

with $0 < S < T$. I will denote the observed values with v_S and v_T respectively (so that $v_S = L_S\widetilde{X}_S + \xi_S$ and $v_T = L_T\widetilde{X}_T + \xi_T$). A general case will follow immediately by induction.

As in van der Meulen and Schauer (2017b, §2.2) define

$$\Upsilon_t := \begin{cases} \mathbb{C}ov\left(\begin{pmatrix}\xi_S \\ \xi_T\end{pmatrix}\right) = \begin{bmatrix}\Sigma_S & 0 \\ 0 & \Sigma_T\end{bmatrix}, & t \in [0,S], \\ \mathbb{C}ov(\xi_T) = \Sigma_T, & t \in (S,T], \end{cases} \qquad \widetilde{L}_t := \begin{cases} \begin{bmatrix} L_S\Phi(S,t) \\ L_T\Phi(T,t)\end{bmatrix}, & t \in [0,S], \\ L_T\Phi(T,t), & t \in (S,T], \end{cases} \tag{6.4}$$

as well as

$$\widetilde{\mu}_t := \int_t^T \widetilde{L}_\tau \widetilde{\beta}_\tau \,\mathrm{d}\tau, \quad t \in [0,T], \qquad \mathrm{x}_{\mathrm{obs}}(t) := \begin{cases} \begin{bmatrix} v_S \\ v_T\end{bmatrix}, & t \in [0,S], \\ v_T, & t \in (S,T]. \end{cases} \tag{6.5}$$

Finally, define a real-valued, time-dependent matrix:

$$\widetilde{M}_t^\dagger := \int_t^T \widetilde{L}_\tau \widetilde{\Gamma}_\tau \widetilde{L}_\tau \,\mathrm{d}\tau + \Upsilon_t, \qquad t \in [0,T], \tag{6.6}$$

and assume:

Assumption A16. *For $t \in [0,T)$ $\widetilde{M}_t^\dagger$ is invertible.*

Remark 6.1.1. The assumption of uniform ellipticity A3 implies assumption A16; however, if A3 does not hold and only A4 does, then assumption A16 might be violated (Bierkens et al., 2018).

Under A16, define:

$$M_t := (M_t^{\dagger})^{-1}. \tag{6.7}$$

With these definition in place it is possible to derive the density for the observations.

Proposition 6.1.1. The density $\widetilde{h}(t,x) := \widetilde{\rho}(\mathtt{x}_{\mathtt{obs}}(t)|x,t)$, $t \in [0,S)$ under the auxiliary law $\widetilde{\mathbb{P}}$ for observing v_S at time S and v_T at time T when the diffusion is conditioned to start from $X_t = x$, $t \in [0,S)$ and the observations are generated via: $v_u = L_u \widetilde{X}_u + \xi_u$, with $\xi_u \sim \mathtt{Gsn}(0,\Sigma_u)$ and $L_u \in \mathbb{R}^{d_u \times d}$, $u \in \{S,T\}$ is given by:

$$\begin{aligned}\widetilde{\rho}(\mathtt{x}_{\mathtt{obs}}(t)|x,t) = (2\pi)^{-(d_S+d_T)/2} |\widetilde{M}_t|^{1/2} \\ \cdot \exp\left\{-\frac{1}{2}\left(\mathtt{x}_{\mathtt{obs}}(t) - \widetilde{\mu}_t - \widetilde{L}_t x\right)^T \widetilde{M}_t \left(\mathtt{x}_{\mathtt{obs}}(t) - \widetilde{\mu}_t - \widetilde{L}_t x\right)\right\}.\end{aligned} \tag{6.8}$$

The density $\widetilde{\rho}(\mathtt{x}_{\mathtt{obs}}(t)|x,t)$, $t \in [S,T]$ can be written in exactly the same way, except for a factor $(2\pi)^{-d_T/2}$ in place of $(2\pi)^{-(d_S+d_T)/2}$.

A density for a similar observational scheme has been derived in van der Meulen and Schauer (2018, proof of lemma 2.5), where the terminal point was assumed to be observed exactly. This case can be recovered with proposition 6.1.1 by setting $L_T = I$ and $\Sigma_T \downarrow 0$.

Simple calculations now reveal that

Proposition 6.1.2. (van der Meulen and Schauer, 2017b, Theorem 2.3), (Bierkens et al., 2018, Lemma 2.5)

$$\widetilde{r}(t,x) = \widetilde{L}_t^T \widetilde{M}_t(\mathtt{x}_{\mathtt{obs}}(t) - \widetilde{\mu}_t - \widetilde{L}_t x), \quad t \in [0,T], \qquad \widetilde{H}(t) := \widetilde{H}_t = \widetilde{L}_t^T \widetilde{M}_t \widetilde{L}_t, \quad t \in [0,T].$$

Therefore, one way of obtaining the expressions

$$(\widetilde{h}(0,x_0), \{\widetilde{r}(t,x), \widetilde{H}(t); (t,x) \in \{(t_0,x_0),\ldots,(t_M,x_M)\}\}),$$

is to simply evaluate the matrix exponentials $\Phi(t)$ at a time-grid $t \in \{t_0,\ldots,t_M\}$ and then perform relevant computations according to eqs. (6.4)–(6.6) and propositions

6.1.1 and 6.1.2. This has been the approach of Schauer et al. (2017), van der Meulen and Schauer (2017a) and van der Meulen and Schauer (2018), however there are two problems with it. First, exponentiating matrices becomes excessively costly as the dimension of the diffusion increases and even for moderate dimensions it is a sub-optimal solution. Second, if a diffusion is observed with noise or if some coordinates are latent, then the Markov property cannot be used to split the problem into independent components, segmented at the times of the observations. It then follows directly from the definitions of $\widetilde{L}_t$, $\widetilde{M}_t$ and $\widetilde{\mu}_t$ that the dimensions of all three elements above increase with the size of the dataset.

To remedy the problem of matrix exponentiation, van der Meulen and Schauer (2017b) show the following

Lemma 6.1.1. (van der Meulen and Schauer, 2017b, Lemma 2.4) $\widetilde{L}$, $\widetilde{M}^\dagger$ and $\widetilde{\mu}$ solve the following backward ordinary differential equations:

$$\begin{cases} \mathrm{d}\widetilde{L}_t = -\widetilde{L}_t \widetilde{B}_t \,\mathrm{d}t, \\ \mathrm{d}\widetilde{M}_t^\dagger = -\widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \,\mathrm{d}t, \\ \mathrm{d}\widetilde{\mu}_t = -\widetilde{L}_t \widetilde{\beta}_t \,\mathrm{d}t, \end{cases} \quad \text{on } t \in (S,T], \text{ with } \begin{cases} \widetilde{L}_T = L_T, \\ \widetilde{M}_T^\dagger = \Sigma_T, \\ \widetilde{\mu}_T = 0, \end{cases} \quad \text{and on } t \in (0,S],$$

$$\text{with } \widetilde{L}_S = \begin{pmatrix} \widetilde{L}_S \\ \widetilde{L}_{S+} \end{pmatrix}, \quad \widetilde{M}_S^\dagger = \begin{bmatrix} \widetilde{\Sigma}_S & 0 \\ 0 & \widetilde{M}_{S+}^\dagger \end{bmatrix}, \quad \widetilde{\mu}_S = \begin{pmatrix} 0 \\ \widetilde{\mu}_{S+} \end{pmatrix},$$

and where $\widetilde{L}_{S+} := \lim_{\epsilon \downarrow 0} \widetilde{L}_{S+\epsilon}$ (with $\widetilde{M}_{S+}^\dagger$ and $\widetilde{\mu}_{S+}$ defined analogously).

This means that matrix exponentials are not necessary for evaluations of $\widetilde{L}$, $\widetilde{M}$ and $\widetilde{\mu}$, and the triplet can instead be approximated by solving the system of ODEs from lemma 6.1.1. Nonetheless, the second issue—of increasing dimension with the size of the dataset—persists. To solve also the second issue van der Meulen and Schauer (2017b) define another set of ordinary differential equations under the assumption:

Assumption A17. *The null-space of* $\widetilde{L}_t$, $t \in [0,T]$ *is equal to* $\{0\}$.

Under A17, it follows directly from proposition 6.1.2 that the process $\widetilde{H}_t^\dagger := (\widetilde{H}_t)^{-1}$, $t \in [0,T]$ exists and as a consequence the following ODEs can be defined

Proposition 6.1.3. (van der Meulen and Schauer, 2017b, Lemma 2.9, 2.10) Let $\widetilde{\nu}_t := \widetilde{H}_t^\dagger \widetilde{L}_t^T \widetilde{M}_t(\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t)$. Then $\widetilde{\nu}_t$ and $\widetilde{H}_t^\dagger$ solve the following backward ODEs

$$\begin{cases} \mathrm{d}\widetilde{H}_t^\dagger = \left[\widetilde{B}_t \widetilde{H}_t^\dagger + \widetilde{H}_t^\dagger \widetilde{B}_t^T - \widetilde{\Gamma}_t\right] \mathrm{d}t, \\ \mathrm{d}\widetilde{\nu}_t = \left[\widetilde{B}_t \widetilde{\nu}_t + \widetilde{\beta}_t\right] \mathrm{d}t, \end{cases} \quad \text{on } t \in (S, T], \text{ with } \begin{cases} \widetilde{H}_T^\dagger = (L_T^T \Sigma^{-1} L_T)^{-1}, \\ \widetilde{\nu}_T = (L_T^T \Sigma^{-1} L_T)^{-1} L_T^T \Sigma_T^{-1} v_T, \end{cases}$$

$$\text{and on } t \in [0, S], \text{ with } \begin{cases} \widetilde{H}_S^\dagger = \widetilde{H}_{S+}^\dagger - \widetilde{H}_{S+}^\dagger L_S^T \left(\Sigma_S + L_S \widetilde{H}_{S+}^\dagger L_S^T\right)^{-1} L_S \widetilde{H}_{S+}^\dagger, \\ \nu_S = \widetilde{H}_S^\dagger \left(L_S^T \Sigma_S^{-1} v_S + \widetilde{H}_{S+} \nu_{S+}\right). \end{cases}$$

Remark 6.1.2. The set of evolution equations defined in proposition 6.1.3 is in fact exactly the Kalman–Bucy filter (Bucy and Joseph, 2005) defined for a backward-evolving linear process $\widetilde{X}$, with no observation noise.

Since by proposition 6.1.2: $\widetilde{r}(t, x) = \widetilde{H}_t(\widetilde{\nu}_t - x)$, so long as assumption A17 holds, $\widetilde{r}$ and $\widetilde{H}$ can be computed on a time grid by solving backward ODEs from proposition 6.1.3. Unlike the solutions to ODEs from lemma 6.1.1, those from proposition 6.1.3 are of constant dimensions in the size of the dataset ($\forall t \in [0, T]$ $\nu_t \in \mathbb{R}^d$ and $\widetilde{H}_t \in \mathbb{R}^{d \times d}$) and thus can be found numerically in a much more efficient way. Nonetheless, they are still sub-optimal for a number of reasons.

First, an extra assumption A17 had to imposed. If it is not satisfied, then the only possibility is to fall back on lemma 6.1.1. Second, a new operation of matrix inverse had to be used—this is not a problem for the computational complexity (as it equals that of matrix multiplication) but it is a source of subtly introduced numerical errors that necessitates very dense (and thus expensive) time grids. Third, perhaps *most importantly*, these ODEs do not help in evaluating $\widetilde{h}(0, x_0)$. Evaluation of $\widetilde{h}(0, x_0)$—though not necessary when the drift and volatility coefficient remain fixed throughout the sampling procedure—is essential for applications to Bayesian inference for diffusion processes and is thus a very important ingredient of the algorithm. Notice from proposition 6.1.1 that at a minimum, $\widetilde{M}_0$ needs to be known to evaluate the normalisation constant of $\widetilde{h}(0, x_0)$.[4] Directly from the definition (i.e. eqs. (6.6) and (6.7)), $\widetilde{M}_0$ is given as an inverse of an integral over the entire time domain of an expression that involves matrix exponentials $\Phi(t)$. Consequently, an approximation to $\widetilde{h}(0, x_0)$ via left-Riemann sums necessitates repeated evaluations of matrix exponentials on a grid of time-points.

[4] In fact, some additional computations reveal that more terms remain unknown

Motivated by these shortcomings, I propose to use an alternative set of variables and the corresponding set of backward ordinary differential equations. First, define the following:

$$\begin{aligned}
\widetilde{F}_t &:= \widetilde{L}_t^T \widetilde{M}_t (\mathrm{x}_{\mathtt{obs}} - \widetilde{\mu}_t), && t \in [0, T], \text{ and} \\
\widetilde{c}_t &:= \frac{1}{2}(\mathrm{x}_{\mathtt{obs}} - \widetilde{\mu}_t)^T \widetilde{M}_t (\mathrm{x}_{\mathtt{obs}} - \widetilde{\mu}_t) + \log\left(\int e^{q_t(v, x_0)} \, \mathrm{d}v \right), && t \in [0, T], \\
& \text{where} \quad q_t(v, x) := -\frac{1}{2}\left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right)^T \widetilde{M}_t \left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right).
\end{aligned} \tag{6.9}$$

It is possible to express $\widetilde{r}$ and $\widetilde{h}$ in terms of $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ as follows:

Proposition 6.1.4. $\widetilde{r}$ and $\log \widetilde{h}(0, x_0)$ admit the following decompositions

$$\widetilde{r}(t, x) = \widetilde{F}_t - \widetilde{H}_t x, \quad t \in [0, T], \qquad \log \widetilde{h}(0, x_0) = -\frac{1}{2}\left\{ x_0^T \widetilde{H}_0 x_0 - 2\widetilde{F}_0^T x_0 \right\} - \widetilde{c}_0.$$

It also turns out that the triplet $(\widetilde{H}, \widetilde{F}, \widetilde{c})$ can be defined as a solution to a system of ordinary differential equations that have a particularly simple form:

Theorem 6.1.1. $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ solve the following ODEs

$$\begin{cases}
\mathrm{d}\widetilde{H}_t = \left(-\widetilde{B}_t^T \widetilde{H}_t - \widetilde{H}_t \widetilde{B}_t + \widetilde{H}_t \widetilde{\Gamma}_t \widetilde{H}_t \right) \mathrm{d}t, \\
\mathrm{d}\widetilde{F}_t = \left(-\widetilde{B}_t^T \widetilde{F}_t + \widetilde{H}_t \widetilde{\Gamma}_t \widetilde{F}_t + \widetilde{H}_t \widetilde{\beta}_t \right) \mathrm{d}t, \qquad \text{on } t \in (S, T], \quad \text{with} \\
\mathrm{d}\widetilde{c}_t = \left(\widetilde{\beta}_t^T \widetilde{F}_t + \frac{1}{2}\widetilde{F}_t^T \widetilde{\Gamma}_t \widetilde{F}_t - \frac{1}{2}\widetilde{H}_t : \widetilde{\Gamma}_t \right) \mathrm{d}t,
\end{cases}$$

$$\widetilde{H}_T = L_T^T \Sigma_T^{-1} L_T, \qquad \widetilde{F}_T = L_T^T \Sigma_T^{-1} v_T \qquad \widetilde{c}_T = \frac{1}{2} v_T^T \Sigma_T^{-1} v_T + \frac{d_T}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_T|,$$

and on $t \in [0, S]$, with $\widetilde{H}_S = \widetilde{H}_{S+} + L_S^T \Sigma_S^{-1} L_S, \qquad \widetilde{F}_S = \widetilde{F}_{S+} + L_S^T \Sigma_S^{-1} v_S$

$$\widetilde{c}_S = \widetilde{c}_{S+} + \frac{1}{2} v_S^T \Sigma_S^{-1} v_S + \frac{d_S}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_S|.$$

There are a couple of remarks to be made about theorem 6.1.1. First, A17 is no longer a prerequisite, so ODEs above can be used to evaluate $\widetilde{H}$ and $\widetilde{r}$ whenever guided proposals are applicable. Second, the dimension of each of $\widetilde{H}_t$, $\widetilde{F}_t$ and $\widetilde{c}_t$ is constant on the entire interval $[0, T]$, regardless of the number of the observations. The only operations that need to be performed are matrix additions and multiplications (and an even simpler Frobenius inner product $A : B := \mathrm{tr}(A^T B)$, for A, B real-valued matrices; inversions of Σ_S and Σ_T have negligible cost, because they do not need to be evaluated on the grid $t_0, \ldots, t_M$ and instead only at the times of the

observations). Third, for numerical purposes, if the observations are made without noise, matrices Σ_S and Σ_T are approximated with ϵI for some small enough ϵ and the inversion is performed accordingly.

The general case of K observations (i.e. when the set of conditioned-on variables is given by (6.1), and not by (6.3)) follows from theorem 6.1.1 by induction

Corollary 6.1.1. $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ solve the same ODEs as in theorem 6.1.1, on $t \in (t_{K-1}, T]$, with

$$\widetilde{H}_T = L_T^T \Sigma_T^{-1} L_T, \qquad \widetilde{F}_T = L_T^T \Sigma_T^{-1} v_T \qquad \widetilde{c}_T = \frac{1}{2} v_T^T \Sigma_T^{-1} v_T + \frac{d_T}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_T|,$$

and on $t \in (t_{i-1}, t_i]$, $(i = K-1, \dots, 1)$, with

$$\widetilde{H}_{t_i} = \widetilde{H}_{t_i+} + L_{t_i}^T \Sigma_{t_i}^{-1} L_{t_i}, \qquad \widetilde{F}_{t_i} = \widetilde{F}_{t_i+} + L_{t_i}^T \Sigma_{t_i}^{-1} v_{t_i}$$
$$\widetilde{c}_{t_i} = \widetilde{c}_{t_i+} + \frac{1}{2} v_{t_i}^T \Sigma_{t_i}^{-1} v_{t_i} + \frac{d_{t_i}}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_{t_i}|, \quad (i = K-1, \dots, 1).$$

The procedure summarising computations of the terms $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ is given in algorithm 6.1 below.

Algorithm 6.1 Solver for $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$

Set $\widetilde{H}_{t_K} \leftarrow L_T^T \Sigma_T^{-1} L_T$
Set $\widetilde{F}_{t_K} \leftarrow L_T^T \Sigma_T^{-1} v_T$
Set $\widetilde{c}_{t_K} \leftarrow \frac{1}{2} v_T^T \Sigma_T^{-1} v_T + \frac{d_T}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_T|$
for $k = K-1, 0$ **do** ▷ Find solutions on dense enough time-grids
    Solve backward ODEs on $t \in (t_k, t_{k+1}]$

$$\begin{cases} \mathrm{d}\widetilde{H}_t = \left(-\widetilde{B}_t^T \widetilde{H}_t - \widetilde{H}_t \widetilde{B}_t + \widetilde{H}_t \widetilde{\Gamma}_t \widetilde{H}_t\right) \mathrm{d}t, \\ \mathrm{d}\widetilde{F}_t = \left(-\widetilde{B}_t^T \widetilde{F}_t + \widetilde{H}_t \widetilde{\Gamma}_t \widetilde{F}_t + \widetilde{H}_t \widetilde{\beta}_t\right) \mathrm{d}t, \\ \mathrm{d}\widetilde{c}_t = \left(\widetilde{\beta}_t^T \widetilde{F}_t + \frac{1}{2} \widetilde{F}_t^T \widetilde{\Gamma}_t \widetilde{F}_t - \frac{1}{2} \widetilde{H}_t : \widetilde{\Gamma}_t\right) \mathrm{d}t, \end{cases}$$

    if $k \neq 0$ **then**
        Set $\widetilde{H}_{t_k} \leftarrow \widetilde{H}_{t_k+} + L_{t_k}^T \Sigma_{t_k}^{-1} L_{t_k}$
        Set $\widetilde{F}_{t_k} \leftarrow \widetilde{F}_{t_k+} + L_{t_k}^T \Sigma_{t_k}^{-1} v_{t_k}$
        Set $\widetilde{c}_{t_k} \leftarrow \widetilde{c}_{t_k+} + \frac{1}{2} v_{t_k}^T \Sigma_{t_k}^{-1} v_{t_k} + \frac{d_{t_k}}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_{t_k}|$

6.2 Re-sampling of the starting point

In the statistics literature it is often assumed for simplicity that the exact value of the starting point x_0 is known (Delyon and Hu, 2006; Bladt et al., 2016). Naturally, this is a somewhat unnatural assumption to be made in practice if all the remaining observations exclude some latent coordinates or are influenced by substantial amount of noise. Thankfully, there is no need to adhere to such artificial constraint when using guided proposals. The procedure for re-sampling the starting point has been described in van der Meulen and Schauer (2017b) under the assumption that $v_0 = L_0 X_0 + \xi_0$, $\xi_0 \sim \texttt{Gsn}(0, \Sigma_0)$ are observed at time 0 and that A17 holds. In this section I will provide a slightly re-formulated version of this statement, which thanks to the results from the previous section need not require assumption A17 to hold.

The main result is as follows:

Theorem 6.2.1. Suppose that X_0 is equipped with a Gaussian prior

$$\pi(X_0) \propto \exp\left\{-\frac{1}{2}(X_0 - \mu_{\text{pr}})^T \Sigma_{\text{pr}}^{-1}(X_0 - \mu_{\text{pr}})\right\}.$$

Then, the posterior density of X_0 under the auxiliary law $\widetilde{\mathbb{P}}$ conditioned on the set of observations v_i, $(i = 1, \ldots, K)$ is Gaussian with mean and covariance:

$$\mu_{\text{post}} := \left(\widetilde{H}_{0+} + \Sigma_{\text{pr}}^{-1}\right)^{-1}\left(\widetilde{F}_{0+} + \Sigma_{\text{pr}}^{-1}\mu_{\text{pr}}\right), \qquad \Sigma_{\text{post}} := \left(\widetilde{H}_{0+} + \Sigma_{\text{pr}}^{-1}\right)^{-1},$$

i.e.

$$\widetilde{\rho}(x_0|\mathcal{Z})\,\mathrm{d}x_0 = (2\pi)^{-d/2}|\Sigma_{\text{post}}|^{-1/2}\exp\left\{-\frac{1}{2}(x_0 - \mu_{\text{post}})^T \Sigma_{\text{post}}^{-1}(x_0 - \mu_{\text{post}})\right\}\mathrm{d}x_0.$$

Remark 6.2.1. Notice that under A17, the statement of van der Meulen and Schauer (2017b, Algorithm 5.1, point 3) follows from theorem 6.2.1, when the prior $\pi(X_0)$ is Gaussian with mean $(L_0^T \Sigma_0^{-1} L_0)^{-1} L_0^T \Sigma_0^{-1} v_0$ and covariance $(L_0^T \Sigma_0^{-1} L_0)^{-1}$.

It is always possible to choose an appropriate prior covariance matrix Σ_{pr} so that $\widetilde{H}_{0+} + \Sigma_{\text{pr}}^{-1}$ is non-singular, which means that the problem of sampling from $\widetilde{\rho}(x_0|\mathcal{Z})$ is well-defined and easy to execute. Nonetheless, I still need to explain how sampling from the posterior over the starting point under the auxiliary law is

supposed to help in handling uncertainty over the starting point under the target law.

To illuminate this connection, let me start from a standard decomposition based on Bayes' rule:

$$\mathbb{P}_b(X \in \cdot|\mathcal{Z}) \propto \pi(X_0)\mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_0)\mathbb{P}_b(X \in \cdot|\mathcal{Z}, X_0).$$

It says that in order to sample a path X under the conditioned target measure $\mathbb{P}_b(\cdot|\mathcal{Z})$ one needs to first sample a starting point according to its posterior under the target law:

$$\mathrm{d}\mathbb{P}_b(X_0|\mathcal{Z}) \propto \pi(X_0)\mathrm{d}\mathbb{P}_b(\mathcal{Z}|X_0),$$

and then, conditionally on the sampled value of X_0, draw the remaining part of the path according to $\mathbb{P}_b(X \in \cdot|\mathcal{Z}, X_0)$. The latter step is exactly the simplified problem of simulating paths under the conditioned target law started at a known, fixed point. For the former step one would ideally wish to sample directly from the posterior under the target law $\mathrm{d}\mathbb{P}_b(X_0|\mathcal{Z})$; however, much like in the case of sampling from $\mathbb{P}_b(X \in \cdot|\mathcal{Z}, X_0)$ this will be impossible and the Metropolis-Hastings step will be indispensable.

This is where theorem 6.2.1 becomes useful. If the auxiliary law is chosen well, then it is reasonable to expect that the density $\widetilde{\rho}(x_0|\mathcal{Z})$ approximates $\mathrm{d}\mathbb{P}_b(x_0|\mathcal{Z})$ faithfully and thus may be used as a proposal density for it. Additionally, because $\widetilde{\rho}(x_0|\mathcal{Z})$ is Gaussian, it is possible to further improve the proposals by exploiting local moves with the preconditioned Crank-Nicolson scheme. To this end, define the centring function for the starting point as

$$\Psi_\theta^{[1]}(Z) := \Sigma_{\mathtt{post}}^{1/2}(\theta)Z + \mu_{\mathtt{post}}(\theta), \tag{6.10}$$

so that if $Z \sim \mathtt{Gsn}(0, I_d)$, then $\Psi_\theta^{[1]}(Z) \sim \mathtt{G}(\mu_{\mathtt{post}}(\theta), \Sigma_{\mathtt{post}}(\theta))$. The sampling is then performed on the space on which the non-centred variable Z is defined. A single update step of the entire path is summarised in algorithm 6.2 below. Note that the update of the path conditioned on the starting point makes use of the preconditioned Crank-Nicolson scheme as well. Its centring function is given by:

$$\Psi_\theta^{[2]}(W, X_0^\circ) := \left\{ X_0^\circ + \int_0^t b_\theta^\circ\left[\Psi_\theta^{[2]}(W, X_0^\circ)_s\right]\mathrm{d}s + \int_0^t \sigma_\theta\left[\Psi_\theta^{[2]}(W, X_0^\circ)_s\right]\mathrm{d}W_s^\circ \,;\, t \in [0, T] \right\}, \tag{6.11}$$

which in practice is determined by an evaluation of the Euler-Maruyama scheme (see example 4.1.2 for details). Function G used below is defined in proposition 3.3.1.

Algorithm 6.2 Non-centrally parametrised path update via guided proposals

Set $\mu_{\texttt{post}} \leftarrow \left(\widetilde{H}_{0+} + \Sigma_{\texttt{pr}}^{-1}\right)^{-1}\left(\widetilde{F}_{0+} + \Sigma_{\texttt{pr}}^{-1}\mu_{\texttt{pr}}\right)$
Set $\Sigma_{\texttt{post}} \leftarrow \left(\widetilde{H}_{0+} + \Sigma_{\texttt{pr}}^{-1}\right)^{-1}$
Draw $Z \sim \texttt{Gsn}(0, I_d)$
Set $Z^\circ \leftarrow \sqrt{\lambda}Z^{(n)} + \sqrt{1-\lambda}Z$
Set $X_0^\circ \leftarrow \Psi_\theta^{[1]}(Z^\circ)$ ▷ See eq. (6.10)
Draw $W \sim \mathbb{W}$ ▷ Law of d'-dimensional Brownian motion on $[0, T]$
Set $W^\circ \leftarrow \sqrt{\lambda}W^{(n)} + \sqrt{1-\lambda}W$
Set $X^\circ \leftarrow \Psi_\theta^{[2]}(W^\circ, X_0^\circ)$ ▷ See eq. (6.11)
Draw $E \sim \texttt{Exp}(1)$
if $E > \log\pi(X^{(n)}) - \log\pi(X_0^\circ) - \int_0^T G(s, X_s^\circ)\,\mathrm{d}s + \int_0^T G(s, X_s^{(n)})\,\mathrm{d}s$ **then**
    Set $(X^{(n+1)}, Z^{(n+1)}, W^{(n+1)}) \leftarrow (X^\circ, Z^\circ, W^\circ)$
else
    Set $(X^{(n+1)}, Z^{(n+1)}, W^{(n+1)}) \leftarrow (X^{(n)}, Z^{(n)}, W^{(n)})$
return $(X^{(n+1)}, Z^{(n+1)}, W^{(n+1)})$

Algorithm 6.2 summarises just a single step of the Metropolis-Hastings algorithm imputing the path. Repeating this step multiple times yields a chain of paths $\{X^{(n)}; n = 0, \ldots, N\}$, whose invariant law is given by $\mathbb{P}_{b_\theta}(\cdot|\mathcal{Z})$. Additionally, algorithm 6.2 can be embedded in the Metropolis-within-Gibbs algorithm that alternately updates parameters and the path (as it was explained chapter 4) so as to perform Bayesian inference. Finally, another way to improve the mixing of the chain $\{X^{(n)}; n = 0, \ldots, N\}$, could be to employ the blocking scheme, similarly to how it was done in van der Meulen and Schauer (2018).

Remark 6.2.2. Let me repeat a very important observation from Clark (1990) that as the times of observations are approached from the left, the significance of the guiding term increases (even up to the point of explosion in the exactly observed setting), so in order to prevent numerical instabilities it is important to keep the discretisation step particularly short for those time-spans. Clark (1990) suggested a certain time-change and scaling aimed at reducing the numerical instabilities and van der Meulen and Schauer (2017a) proposed its further extension. For the setting considered in this thesis, the joint space-time transformation cannot be applied in

full generality (this happens for instance when some coordinates remain latent). Consequently, I resort to a partial transformation of the one presented in van der Meulen and Schauer (2017a): i.e. I set up an equidistant time-grid and then transform it via: $\tau_i(s):[t_{i-1},t_i]\to[t_{i-1},t_i]$ given by $\tau_i(s)=t_{i-1}+s(2-s/(t_i-t_{i-1}))$, $(i=1,\ldots,K)$. I used this transformation in the numerical examples in the subsequent section.

Additionally, let me remark that the ordinary differential equations for $\widetilde{L}$, $\widetilde{M}^\dagger$ and $\widetilde{\mu}$ hold one advantage over those for $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ and that is that they numerically more stable. The difference between the two is negligible for most regimes that I have tested—especially after employing a high level numerical solver (I used 7th order, 7/6 Runge-Kutta method due to Verner (1978))—except for the time directly adjacent to the time of exact and full observations of the process (exact, but partial observations are not afflicted). At those times the "zero-level noise" is approximated with a Gaussian noise, whose covariance matrix is ϵI with ϵ *very small*. Employing $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ in such settings either calls for very dense time-grids or use of an insufficiently small ϵ. Fortunately, an easy to implement and a very successful remedy to this problem is to simply employ the solvers for $\widetilde{L}$, $\widetilde{M}^\dagger$ and $\widetilde{\mu}$ on the short time segments directly adjacent to the times of the exact observations (for which $\widetilde{L}$, $\widetilde{M}^\dagger$ and $\widetilde{\mu}$ have a fixed and small dimension) and use $\widetilde{H}$, $\widetilde{F}$ and $\widetilde{c}$ for all the remaining times.

6.3 Numerical results

In this section I apply guided proposals to the problem of parameter inference for a stochastic version of the Lorenz system (Lorenz, 1963). At the time of its conception, this system of ODEs has been developed to model atmospheric convection; however, since then it has found applications across engineering and other sciences (Sparrow, 2012). The system is perhaps best known for its chaotic behaviour, with trajectories resembling (under certain choices of parametrisations) the shape of a butterfly. The problem of inference for a stochastic version of this system is well-known for its difficulty. Additionally, the drift is not of the potential form (i.e. assumption A7 does not hold), so a number of competing, Bayesian inference methods cannot even be applied. As I demonstrate below, even when this problem is

made much more difficult by recording only partial and noisy observations of the process, guided proposals are successful in recovering the posterior over the parameters.

6.3.1 Stochastic Lorenz system

The stochastic version of the Lorenz system is defined by the following three dimensional stochastic differential equation:

$$\begin{gathered} \mathrm{d}X_t = b_\theta(X_t)\,\mathrm{d}t + \theta^{[4]} I_3\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \\ \text{where} \quad b_\theta(x) := \begin{pmatrix} -\theta^{[1]}(x^{[2]} - x^{[1]}) \\ x^{[1]}(\theta^{[2]} - x^{[3]}) - x^{[2]} \\ x^{[1]} x^{[2]} - \theta^{[3]} x^{[3]} \end{pmatrix}, \end{gathered} \tag{6.12}$$

$\theta^{[1:3]} \in \mathbb{R}^3$, $\theta^{[4]} \in \mathbb{R}_+$ are some parameters of interest, I_3 denotes a 3×3 identity matrix and W is a 3-dimensional Brownian motion.

For the experiments I simulated $K = 20$ equidistant observations on the interval $[0, T]$, with $T = 4$ (i.e. $t_i = 0.2i$, $i = 1, \ldots, K$). The first coordinate was latent whereas the trailing two were disturbed by addition of some Gaussian noise, centred at 0 and with covariance matrix I_2. Consequently, the observational scheme can be encoded with:

$$L_i = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \Sigma_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad i = 1, \ldots, K.$$

The full trajectory from which the observations were recorded is plotted in fig. 6.1. Notice that between each observation the process follows highly non-linear trajectories. Removing all information about the location of the first coordinate and disturbing the remaining two with some Gaussian noise adds on an extra layer of difficulty to this problem.

6.3.1.1 Auxiliary law for guided proposals

The auxiliary law for guided proposals was initially chosen to be induced by the stochastic differential equations defined separately on each sub-interval $[t_{i-1}, t_i]$, $i = 1, \ldots, K$ by linearising the Lorenz system at the terminal observation of this

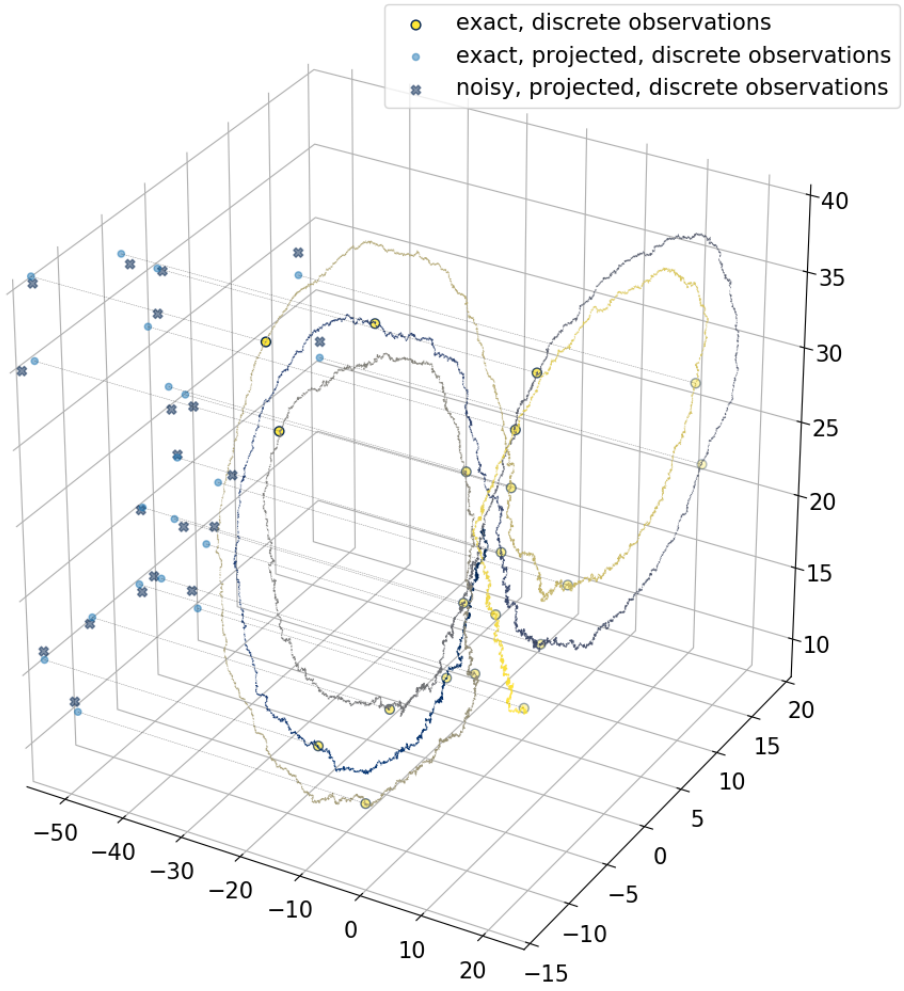


Figure 6.1: Observations of the stochastic Lorenz system. The entire (unobserved) trajectory of the three dimensional process X is plotted with a solid line that changes colour from dark blue to light yellow so as to indicate the temporal component. Exact, discrete observations, recorded at times $t_i = 0.2i$, $i = 1, \ldots, K$ are marked with two-coloured dots, though they are not observed. Their projections onto the YZ plane are marked with blue dots, these are not observed either. The noisy versions of the projected observations are marked with crosses and these are the observations for the experiments.

interval, i.e.:

$$\mathrm{d}\widetilde{X}_t = \left(B_\theta[\tilde{x}^{[1]}]\widetilde{X}_t + \beta_\theta[\tilde{x}^{[1]}]\right)\mathrm{d}t + \theta^{[4]} I_3 \,\mathrm{d}W_t, \quad t \in [t_{i-1}, t_i],$$
$$\text{where} \quad B_\theta[x] := \begin{pmatrix} -\theta^{[1]} & \theta^{[1]} & 0 \\ \theta^{[2]} - v_i^{[3]} & -1 & -x \\ v_i^{[2]} & x & -\theta^{[3]} \end{pmatrix}, \quad \text{and} \quad \beta_\theta[x] := \begin{pmatrix} 0 \\ x v_i^{[3]} \\ -x v_i^{[2]} \end{pmatrix}, \tag{6.13}$$

where $v_i^{[2:3]}$ is the observation recorded at the time t_i and where $\bar{x}^{[1]} \in \mathbb{R}$ is some arbitrary constant that ideally should be equal to the unobserved value of the first coordinate at the time t_i. In practice, the algorithm is robust to the choice of $\bar{x}^{[1]}$, because as I explain below, it is possible to tune the auxiliary law in an adaptive way.

6.3.1.2 Adapting the auxiliary law

As suggested in van der Meulen and Schauer (2017b) in the setting of path-sampling, it is possible to adaptively tune the auxiliary law so as to improve the quality of the proposal paths. Owing to the results presented in this chapter, this approach can now also extended to the setting of statistical inference.

The idea is based on a technique from Whitaker et al. (2017) for approximating non-linear diffusions with linear ones. In particular, to approximate a non-linear diffusion:

$$\mathrm{d}V_t = \alpha(V_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad V_0 = v_0, \quad t \in [0, T],$$

where σ is some constant volatility matrix, one could use a linear diffusion whose drift is obtained from Taylor expansion of α around the mean trajectory $\bar{x}_t := \mathbb{E}[X_t|\mathscr{D}]$, i.e. it is given by:

$$\widetilde{\alpha}(t, x) := \widetilde{\alpha}(\bar{x}_t) + \mathscr{J}_\alpha(\bar{x}_t)(x - \bar{x}_t),$$

where $\mathscr{J}_\alpha$ denotes the Jacobian matrix of α.

Consequently, if $\bar{X}_t^{(\theta)} := \mathbb{E}[X_t|\theta, \mathscr{D}]$ were known for each parameter θ, then in the example of the Lorenz system the following process could be chosen to induce the auxiliary law:

$$\widetilde{X}_t = \left(B_\theta^{\mathtt{adpt}}[\bar{X}_t^{(\theta)}]\widetilde{X}_t + \beta_\theta^{\mathtt{adpt}}[\bar{X}_t^{(\theta)}]\right)\mathrm{d}t + \theta^{[4]} I_3\,\mathrm{d}W_t, \quad t \in [t_{i-1}, t_i], \quad \text{where}$$
$$B_\theta^{\mathtt{adpt}}[x] := \begin{pmatrix} -\theta^{[1]} & \theta^{[1]} & 0 \\ \theta^{[2]} - x^{[3]} & -1 & -x^{[1]} \\ x^{[2]} & x^{[1]} & -\theta^{[3]} \end{pmatrix}, \quad \text{and} \quad \beta_\theta^{\mathtt{adpt}}[x] := b_\theta(x) - B_\theta^{\mathtt{adpt}}[x]x, \tag{6.14}$$

with $b_\theta(x)$ defined in eq. (6.12). Of course, in practice, $\bar{X}_t^{(\theta)}$ is unknown, but during the initial stages of running an MCMC sampler it is possible to learn an approximation $\bar{X}_t \approx \mathbb{E}_\theta[\mathbb{E}[X_t|\theta, \mathscr{D}]|\mathscr{D}]$ and use it in eq. (6.14) in place of $\bar{X}_t^{(\theta)}$. To avoid

common pitfalls with adaptive schemes, the number of times that the auxiliary law is adapted needs to be fixed a-priori. Additionally, departure from the initial choice of the auxiliary law can be done incrementally, by using the following process $\widetilde{X}$:

$$\begin{aligned} \mathrm{d}\widetilde{X}_t = \Big\{\Big(\rho B_\theta[\bar{X}^{[1]}_{t_i}] + (1-\rho)B^{\mathtt{adpt}}_\theta[\bar{X}_t]\Big)\widetilde{X}_t + \rho\beta_\theta[\bar{X}^{[1]}_{t_i}] + (1-\rho)\beta^{\mathtt{adpt}}_\theta[\bar{X}_t]\Big\}\,\mathrm{d}t \\ + \theta^{[4]} I_3\,\mathrm{d}W_t, \quad \widetilde{X}_0 = x_0, \quad t \in [t_{i-1}, t_i], \end{aligned} \tag{6.15}$$

where ρ is gradually decreased from 1 to $a \geq 0$, for some user-chosen constant a and where $\bar{X}^{[1]}_{t_i}$ is initially chosen to be equal to some arbitrary constant $\bar{x}^{[1]} \in \mathbb{R}$.

6.3.1.3 Inference results

For the results presented below I used auxiliary law given in eq. (6.15), initially with $\rho = 1$ and used preconditioned Crank-Nicolson scheme with memory parameter $\lambda = 0.92$. After 500 iterations I computed the mean trajectory $\bar{X}$ using the accepted paths of the Markov chain and I updated the auxiliary law according to eq. (6.14), setting $\rho = 0.7$. After further 500 iterations I re-computed the mean trajectory $\bar{X}$ using the most recent 500 paths and updated the auxiliary law according to eq. (6.14), setting $\rho = 0.4$. The memory parameter was updated to $\lambda = 0.9$. The adaptation was repeated 3 more times, each time based on the most recent 500 iterations, with ρ set to 0.2 each time and the memory parameter being decreased to 0.88, 0.83 and finally 0.8. No further adaptations were performed beyond that point (i.e. after 2500th iteration). The results of the inference are given in fig. 6.2. The joint posterior over all four parameters has been successfully identified. The volatility parameter is particularly difficult to identify in this setting because of two reasons: first, the observations are rather sparse and the number of observations low, which means that the data carries relatively little information about the quadratic variation; second, the observation noise is of a similar order as the magnitude of the volatility, which further clouds information about this parameter.

6.4 Discussion

In this chapter I re-formulated some of the core computational routines of guided proposals and this in turn resulted in two improvements: one, in a reduction of

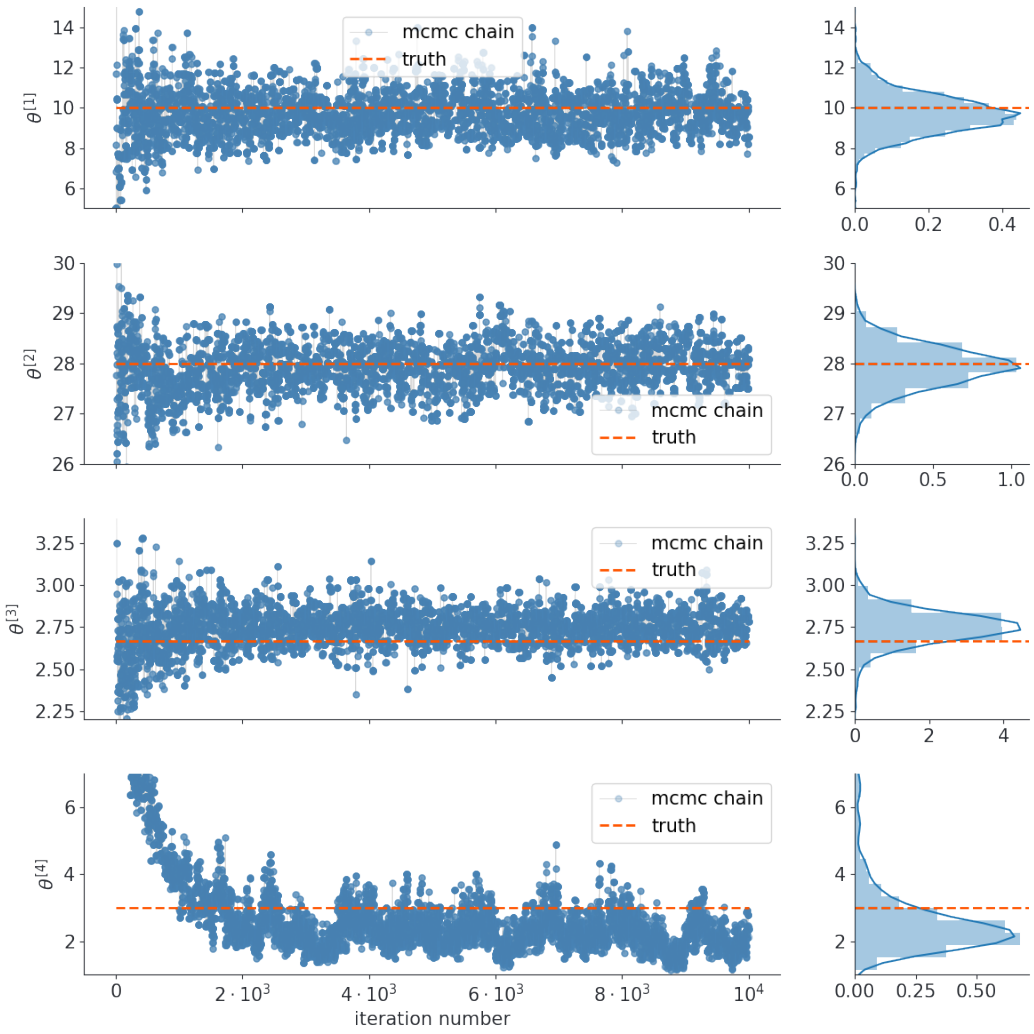


Figure 6.2: Inference results for the stochastic Lorenz system. The chain was initialised at $\theta^{(0)} = (5, 15, 6, 8)$.

the computational cost of the algorithm, and two, in a possibility of automating statistical inference with guided proposals through encapsulation of the vast majority of the necessary computer code within a generic implementation. Only the details pertinent to any particular example need to be coded by a practitioner. The novel formulation was achieved by deriving a set of ordinary differential equations (whose solutions are constant in the dimension of the dataset), which can be used to compute all the non-trivial, unknown terms required for an implementation of guided proposals.

In the numerical section I illustrated that thanks to a generic way in which computations are performed, the auxiliary law may be chosen very fittingly to the

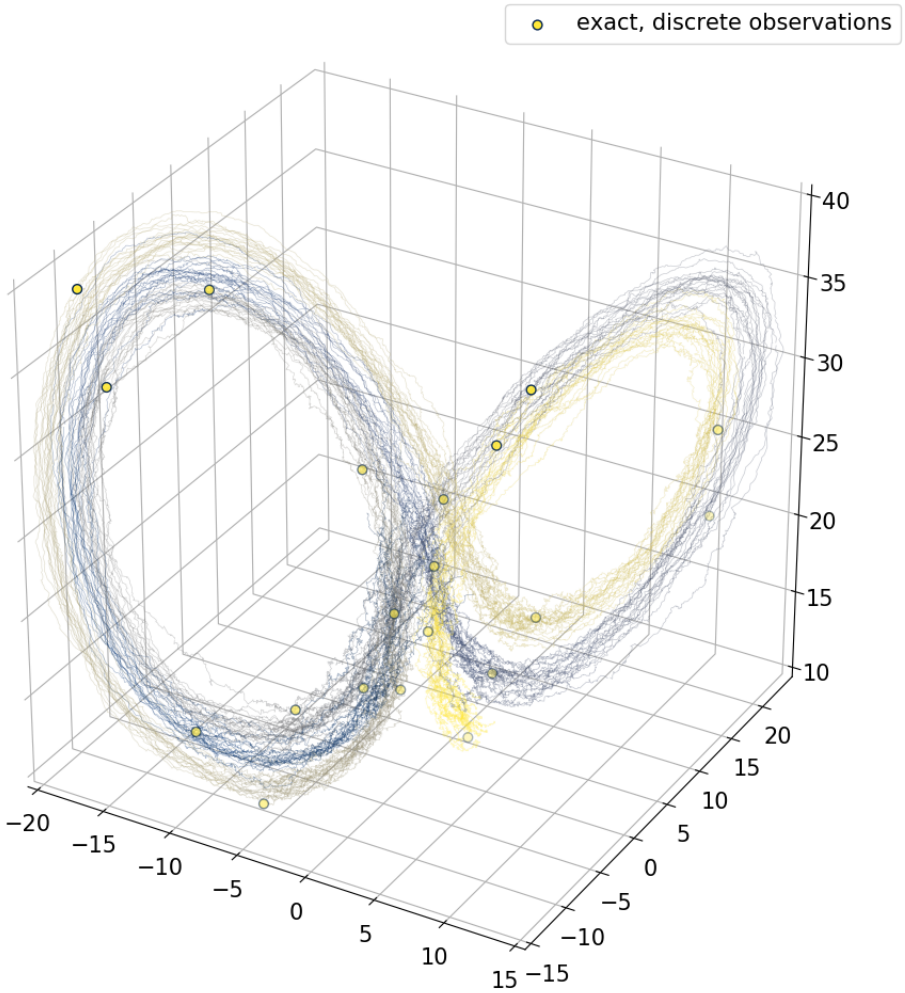


Figure 6.3: Paths of the Lorenz system simulated during inference. The 20 plotted paths were obtained as a result of thinning the chain of trajectories that were drawn during inference from the iterations numbered $8 \cdot 10^3$–10^4. Recall that only partially and noisy observations were recorded, the exact observations are plotted merely as a visual aid.

task and as a result I showed that guided proposals may achieve high performance on challenging inference problems.

Nonetheless, in some sense this is but a modest display of what this method may be capable of. All the computations have been performed on a single thread of a mid-range, customer-grade, 6th generation Intel CPU (i7-6850K CPU) in less than 15 minutes. Since nearly all of the computational resources are invested into matrix additions and multiplications one can move nearly all of these computations onto a GPU and run the code on a newer machine to tackle high-dimensional diffusion

problems, without resorting to any improvements to the methodology.

However, even the progress on the methodology itself is possible and this the subject of an ongoing research. The path for some of those enhancements is quite clear. First, as written by F. van der Meulen in the `Bridge.jl` package (and soon to appear in a publication), it is possible to leverage automatic differentiation libraries so as to derive gradients of the log-likelihood function with respect to the value of the starting point and implement updates of this point via Hamiltonian Monte Carlo or Metropolis-adjusted Langevin algorithm. The same extension can be implemented for the updates of the parameters. Additionally, many high-dimensional diffusion problems (such as, say, finite-dimensional approximations to stochastic partial differential equations) have an element of sparsity to them, and this may be leveraged by a number of computations—from solving the system of ODEs, to sampling the proposal paths. All of the above and more to come will be added on a continual basis to the `BridgeSDEInferece.jl` package.

Proofs

Proof of proposition 6.1.1. Notice that $\mathcal{V} := (V_S^T, V_T^T)^T := ((L_S\widetilde{X}_S + \xi_S)^T, (L_T\widetilde{X}_T + \xi_T)^T)^T$ must be Gaussian, as it is constructed from a linear combination of Gaussian vectors. Its mean and covariance follow easily from the expressions in eq. (6.2):

$$\bar{v} := \begin{pmatrix} L_S\mu_S(t,x) \\ L_T\mu_T(t,x) \end{pmatrix}, \quad \bar{\Omega} := \begin{pmatrix} L_SK_{SS}L_S^T + \Sigma_S & L_SK_{ST}L_T^T \\ L_TK_{TS}L_S^T & L_TK_{TT}L_T^T + \Sigma_T \end{pmatrix},$$

with $K_{\tau\nu} := K_{\tau\wedge\nu}(t)$, $\tau, \nu \in [t, T]$. Careful comparison of $\bar{v}$ and $\bar{\Omega}$ with the definitions of $\widetilde{M}_t^\dagger$, $\widetilde{L}_t$ and $\widetilde{\mu}_t$ reveals that

$$\bar{v} = \widetilde{L}_t x + \widetilde{\mu}_t, \qquad \bar{\Omega} = \widetilde{M}_t^\dagger,$$

and the result follows (the density under the assumption $t \in [S, T]$ follows analogously). □

Proof of proposition 6.1.4. The expression for $\widetilde{r}$ follows immediately from proposition 6.1.2 and the definition of $\widetilde{F}$. To derive the expression for $\widetilde{h}$, I start from eq. (6.8). Notice that (using the notation from eq. (6.9)):

$$\widetilde{h}(t,x) = \widetilde{\rho}(\mathrm{x}_{\mathtt{obs}}(t)|x,t) = \frac{1}{\mathscr{Z}} \exp\{q_t(\mathrm{x}_{\mathtt{obs}}, x)\}, \quad \text{where } \mathscr{Z} := (2\pi)^{(d_S+d_T)/2}|\widetilde{M}_t|^{-1/2}. \tag{6.16}$$

$\mathscr{Z}$ is a normalisation constant and thus it can also be written as:

$$\mathscr{Z} = \int \exp\{q_t(v,x)\}\,\mathrm{d}v.$$

Derivation of $\log\widetilde{h}$ now follows by simple algebra and the definitions of $\widetilde{F}$, $\widetilde{H}$ and $\widetilde{c}$:

$$\begin{aligned}
\log\widetilde{h}(t,x) &= q_t(\mathtt{x}_{\mathtt{obs}}(t),x) - \log\mathscr{Z} \\
&= -\frac{1}{2}(\mathtt{x}_{\mathtt{obs}}(t)-\widetilde{\mu}_t)^T\widetilde{M}_t(\mathtt{x}_{\mathtt{obs}}(t)-\widetilde{\mu}_t) - \log\int e^{q_t(v,x)}\,\mathrm{d}v \\
&\quad -\frac{1}{2}\Big(-2x^T\widetilde{L}_t^T\widetilde{M}_t(\mathtt{x}_{\mathtt{obs}}(t)-\widetilde{\mu}_t) + x^T\widetilde{L}_t^T\widetilde{M}_t\widetilde{L}_t x\Big) \\
&= -\widetilde{c}_t - \frac{1}{2}\Big(-2\widetilde{F}_t^T x + x^T\widetilde{H}_t x\Big).
\end{aligned}$$

Setting $t = 0$ and $x = x_0$ in the expression above yields the result. □

Proof of theorem 6.1.1. The proof is split into two parts. In the first one I derive the set of backward ordinary differential equations solved by $\widetilde{H}_t$, $\widetilde{F}_t$ and $\widetilde{c}_t$. In the second part I show how to compute the *update equations* giving values of $\widetilde{H}_u$, $\widetilde{F}_u$ and $\widetilde{c}_u$, $u \in \{S,T\}$ at the times of the observations.

Backward ordinary differential equations An ordinary differential equation solved by $\widetilde{H}_t$ has been derived as a by-product in the proof of van der Meulen and Schauer (2017b, Lemma 2.9). For completeness, I re-state the relevant part of this proof in eqs. (6.17)–(6.19):

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{H}_t &= \Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{L}_t\Big)^T\widetilde{M}_t\widetilde{L}_t + \widetilde{L}_t^T\Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t\Big)\widetilde{L}_t + \widetilde{L}_t^T\widetilde{M}_t\Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{L}_t\Big) \\
&= -\widetilde{B}_t^T\widetilde{L}_t^T\widetilde{M}_t\widetilde{L}_t + \widetilde{L}_t^T\Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t\Big)\widetilde{L}_t - \widetilde{L}_t^T\widetilde{M}_t\widetilde{L}_t\widetilde{B}_t \\
&\quad -\widetilde{B}_t^T\widetilde{H}_t - \widetilde{H}_t\widetilde{B}_t + \widetilde{L}_t^T\Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t\Big)\widetilde{L}_t.
\end{aligned} \tag{6.17}$$

Since $\widetilde{M}_t := (\widetilde{M}_t^\dagger)^{-1}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t = -\widetilde{M}_t\Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t^\dagger\Big)\widetilde{M}_t = \widetilde{M}_t\widetilde{L}_t\widetilde{\Gamma}_t\widetilde{L}_t^T\widetilde{M}_t, \tag{6.18}$$

and thus:

$$\widetilde{L}_t^T\Big(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t\Big)\widetilde{L}_t = \widetilde{H}_t\widetilde{\Gamma}_t\widetilde{H}_t. \tag{6.19}$$

Substituting this back into eq. (6.17) yields:

$$\mathrm{d}\widetilde{H}_t = \left(-\widetilde{B}_t^T \widetilde{H}_t - \widetilde{H}_t \widetilde{B}_t + \widetilde{H}_t \widetilde{\Gamma}_t \widetilde{H}_t\right) \mathrm{d}t$$

For $\widetilde{F}_t$ notice:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{F}_t &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{L}_t\right)^T \widetilde{M}_t(\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) + \widetilde{L}_t^T \left(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t\right)(\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) - \widetilde{L}_t^T \widetilde{M}_t \left(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mu}_t\right) \\
&= -\widetilde{B}_t^T \widetilde{L}_t^T \widetilde{M}_t (\mathrm{x}_{\mathsf{obs}} - \widetilde{\mu}_t) + \widetilde{L}_t^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t (\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) + \widetilde{L}_t^T \widetilde{M}_t \widetilde{L}_t \widetilde{\beta}_t \\
&= -\widetilde{B}_t^T \widetilde{F}_t + \widetilde{H}_t \widetilde{\Gamma}_t \widetilde{F}_t + \widetilde{H}_t \widetilde{\beta}_t.
\end{aligned}$$

To derive an ODE for $\widetilde{c}$ I first split $\widetilde{c}$ into two terms:

$$\begin{gathered}
\widetilde{c}_t := \frac{1}{2} C_t^{(1)} + C_t^{(2)}, \quad \text{where} \\
C_t^{(1)} := (\mathrm{x}_{\mathsf{obs}} - \widetilde{\mu}_t)^T \widetilde{M}_t (\mathrm{x}_{\mathsf{obs}} - \widetilde{\mu}_t) \quad \text{and} \quad C_t^{(2)} := \log\left(\int e^{q_t(v, x_0)} \mathrm{d}v\right).
\end{gathered} \tag{6.20}$$

An ODE for the first term follows by simple algebra, eq. (6.18) and a definition of $\widetilde{F}_t$ in eq. (6.9)

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} C_t^{(1)} &= -2\left(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\mu}_t\right)^T \widetilde{M}_t(\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) + (\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t)^T \left(\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{M}_t\right)(\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) \\
&= 2\widetilde{\beta}_t^T \widetilde{L}_t^T \widetilde{M}_t (\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) + (\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t)^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t (\mathrm{x}_{\mathsf{obs}}(t) - \widetilde{\mu}_t) \\
&= 2\widetilde{\beta}_t^T \widetilde{F}_t + \widetilde{F}_t^T \widetilde{\Gamma}_t \widetilde{F}_t.
\end{aligned} \tag{6.21}$$

To find an ordinary differential equation solved by $C_t^{(2)}$ observe that:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} C_t^{(2)} &= e^{-C_t^{(2)}} \int e^{q_t(v,x)} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\left\{-\frac{1}{2}\left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right)^T \widetilde{M}_t \left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right)\right\} \mathrm{d}v \\
&= -\frac{1}{2} e^{-C_t^{(2)}} \int e^{q_t(v,x)} \left\{2\left(\widetilde{L}_t \widetilde{\beta}_t + \widetilde{L}_t \widetilde{B}_t x\right)^T \widetilde{M}_t \left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right)\right. \\
&\qquad \left. + \left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t \left(v - \widetilde{\mu}_t - \widetilde{L}_t x\right)\right\} \mathrm{d}v \\
&= -\frac{1}{2}\left\{2\left(\widetilde{\beta}_t + \widetilde{B}_t x\right)^T \widetilde{L}_t^T \widetilde{M}_t \left(-\widetilde{\mu}_t - \widetilde{L}_t x\right)\right. \\
&\qquad \left. + \left(\widetilde{\mu}_t + \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t \left(\widetilde{\mu}_t + \widetilde{L}_t x\right)\right\} e^{-C_t^{(2)}} \int e^{q_t(v,x)} \mathrm{d}v \\
&\quad -\frac{1}{2} \cdot 2\left\{\left(\widetilde{\beta}_t + \widetilde{B}_t x\right)^T \widetilde{L}_t^T \widetilde{M}_t - \left(\widetilde{\mu}_t + \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t\right\} e^{-C_t^{(2)}} \int v e^{q_t(v,x)} \mathrm{d}v \\
&\quad -\frac{1}{2} e^{-C_t^{(2)}} \int v^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t v e^{q_t(v,x)} \mathrm{d}v.
\end{aligned} \tag{6.22}$$

where I used the Leibniz rule to exchange integration and differentiation, definitions of $\widetilde{\mu}_t$ and $\widetilde{L}_t$ (given in eqs. (6.4) and (6.5)) to express their derivatives with respect to t and eq. (6.18). Since by eq. (6.16):

$$\widetilde{h}(t,x) = e^{-\mathrm{C}_t^{(2)}} e^{q_t(\mathtt{x}_{\mathtt{obs}}(t),x)},$$

is the density for the observations $\mathtt{x}_{\mathtt{obs}}(t)$, which is Gaussian with mean $\widetilde{\mu}_t + \widetilde{L}_t x$ and covariance matrix $\widetilde{M}_t^{\dagger}$, I have

$$e^{-\mathrm{C}_t^{(2)}} \int e^{q_t(v,x)}\,\mathrm{d}v = 1, \quad \text{and} \quad e^{-\mathrm{C}_t^{(2)}} \int v e^{q_t(v,x)}\,\mathrm{d}v = \widetilde{\mu}_t + \widetilde{L}_t x.$$

Consequently, the first two terms in the final expression of eq. (6.22) simplify to:

$$\frac{1}{2}\left(\mu_t + \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{a}_t \widetilde{L}_t^T \widetilde{M}_t \left(\mu_t + \widetilde{L}_t x\right). \tag{6.23}$$

The third and final term in the rightmost expression of eq. (6.22) multiplied by -2 expands to:

$$\begin{aligned} e^{-\mathrm{C}_t^{(2)}} \int v^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t v e^{q_t(v,x)}\,\mathrm{d}v &= \mathrm{tr}\left(\widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t \widetilde{M}_t^{-1}\right) \\ &\quad + \left(\widetilde{\mu}_t + \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t \left(\widetilde{\mu}_t + \widetilde{L}_t x\right) \\ &= \mathrm{tr}\left(\widetilde{L}_t^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t\right) + \left(\widetilde{\mu}_t + \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{\Gamma}_t \widetilde{L}_t^T \widetilde{M}_t \left(\widetilde{\mu}_t + \widetilde{L}_t x\right) \\ &= \mathrm{tr}(\widetilde{H}_t \widetilde{\Gamma}_t) + \left(\widetilde{\mu}_t + \widetilde{L}_t x\right)^T \widetilde{M}_t \widetilde{L}_t \widetilde{a}_t \widetilde{L}_t^T \widetilde{M}_t \left(\widetilde{\mu}_t + \widetilde{L}_t x\right), \end{aligned} \tag{6.24}$$

where I used that the expected value of a quadratic form $Y^T A Y$, with a symmetric $A \in \mathbb{R}^{d\times d}$ and Y denoting a random variable with mean μ and covariance Σ, is given by:

$$\mathbb{E}[Y^T A Y] = \mathrm{tr}(A\Sigma) + \mu^T A \mu.$$

Finally, by (6.22), $\frac{\mathrm{d}}{\mathrm{d}t} C_t^{(2)}$ is given by the sum of (6.23) and a negative one half of (6.24), which yields:

$$\frac{\mathrm{d}}{\mathrm{d}t} C_t^{(2)} = -\frac{1}{2}\mathrm{tr}(\widetilde{H}_t \widetilde{\Gamma}_t) = -\frac{1}{2} \widetilde{H}_t : \widetilde{\Gamma}_t \tag{6.25}$$

The decomposition in eq. (6.20), together with the expressions in eqs. (6.21) and (6.25) yield:

$$\mathrm{d}\widetilde{c}_t = \left(\widetilde{\beta}_t^T \widetilde{F}_t + \frac{1}{2}\widetilde{F}_t^T \widetilde{\Gamma}_t \widetilde{F}_t - \frac{1}{2}\widetilde{H}_t : \widetilde{\Gamma}_t\right) \mathrm{d}t.$$

Update equations The boundary conditions for $\widetilde{H}_T$, $\widetilde{F}_T$ and $C_T^{(1)}$:

$$\widetilde{H}_T = L_T^T \Sigma_T^{-1} L_T, \quad \widetilde{F}_T = L_T^T \Sigma_T^{-1} v_T, \quad C_T^{(1)} = v_T \Sigma_T^{-1} v_T,$$

follow directly from the definitions: $\widetilde{L}_T := L_T$, $\widetilde{M}_T := \Sigma_T^{-1}$ and $\widetilde{\mu}_T := 0$; as well as the expressions for $\widetilde{H}_t$, $\widetilde{F}_t$ and $C_t^{(1)}$ given in proposition 6.1.2, eq. (6.9) and eq. (6.20) respectively. The boundary condition for $C_T^{(2)}$ follows from the definition eq. (6.20), and the fact that $\int e^{q_t(v,x)} \mathrm{d}v$ is a normalisation constant $\mathscr{Z}$ from (6.16), and thus:

$$C_T^{(2)} = \frac{d_T}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_T|.$$

Consequently, the boundary condition for $\widetilde{c}_T$ is given by

$$\widetilde{c}_T = \frac{1}{2} C_T^{(1)} + C_T^{(2)} = \frac{1}{2} v_T \Sigma_T^{-1} v_T + \frac{d_T}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma_T|.$$

To derive the update equations at time S, recall from lemma 6.1.1 the update equations for $\widetilde{L}_S$, $\widetilde{M}_S$ and $\widetilde{\mu}_S$:

$$\widetilde{L}_S = \begin{pmatrix} \widetilde{L}_S \\ \widetilde{L}_{S+} \end{pmatrix}, \quad \widetilde{M}_S = \begin{bmatrix} \widetilde{\Sigma}_S & 0 \\ 0 & \widetilde{M}_{S+}^{\dagger} \end{bmatrix}, \quad \widetilde{\mu}_S = \begin{pmatrix} 0 \\ \widetilde{\mu}_{S+} \end{pmatrix}.$$

Combining the expressions for $\widetilde{H}_t$, $\widetilde{F}_t$ and $C_t^{(1)}$ given in proposition 6.1.2, eq. (6.9) and eq. (6.20) respectively, with the update equations above yields the update equations for $\widetilde{H}_S$, $\widetilde{F}_S$ and $C_S^{(1)}$:

$$\widetilde{H}_S = \widetilde{H}_{S+} + L_S^T \Sigma_S^{-1} L_S, \quad \widetilde{F}_S = \widetilde{F}_{S+} + L_S^T \Sigma_S^{-1} v_S, \quad C_S^{(1)} = C_{S+}^{(1)} + v_S \Sigma_S^{-1} v_S.$$

To derive the update equations for $C_S^{(2)}$, notice first:

$$\begin{aligned} q_S(v,x) &= -\frac{1}{2} \left[\begin{pmatrix} v_S \\ \mathtt{x}_{\mathtt{obs}}(S+) \end{pmatrix} - \begin{pmatrix} 0 \\ \widetilde{\mu}_{S+} \end{pmatrix} - \begin{pmatrix} L_S \\ \widetilde{L}_{S+} \end{pmatrix} x \right]^T \begin{pmatrix} \Sigma_S^{-1} & 0 \\ 0 & \widetilde{M}_{S+} \end{pmatrix} \\ &\qquad \cdot \left[\begin{pmatrix} v_S \\ \mathtt{x}_{\mathtt{obs}}(S+) \end{pmatrix} - \begin{pmatrix} 0 \\ \widetilde{\mu}_{S+} \end{pmatrix} - \begin{pmatrix} L_S \\ \widetilde{L}_{S+} \end{pmatrix} x \right] \\ &= -\frac{1}{2} (v_S - L_S x)^T \Sigma_S^{-1} (v_S - L_S x) + q_{S+}(\mathtt{x}_{\mathtt{obs}}(S+), x). \end{aligned}$$

I can now write:

$$\begin{aligned}
C_S^{(2)} &:= \log\left(\int e^{q_S(v,x)}\,\mathrm{d}v\right) \\
&= \log\left(\int e^{-\frac{1}{2}(v_S - L_S x)^T \Sigma_S^{-1}(v_S - L_S x) + q_{S+}(\mathtt{x}_{\mathtt{obs}}(S+),x)}\,\mathrm{d}v_S\,\mathrm{d}(\mathtt{x}_{\mathtt{obs}}(S+))\right) \\
&= C_{S+}^{(2)} + \log\left(\int e^{-\frac{1}{2}(v_S - L_S x)^T \Sigma_S^{-1}(v_S - L_S x)}\,\mathrm{d}v_S\right) \\
&= C_{S+}^{(2)} + \frac{d_S}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_S|,
\end{aligned}$$

where I used that $\int e^{-\frac{1}{2}(v_S - L_S x)^T \Sigma_S^{-1}(v_S - L_S x)}\,\mathrm{d}v_S$ is the normalisation constant of a (d_S-dimensional) Gaussian density with mean $L_S x$ and covariance Σ_S. The update equation for $\widetilde{c}_S$ follows readily

$$\widetilde{c}_S = \frac{1}{2}C_S^{(1)} + C_S^{(2)} = \widetilde{c}_{S+} + v_S \Sigma_S^{-1} v_S + \frac{d_S}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_S|.$$

□

Proof of theorem 6.2.1. Assume for simplicity that $\mathcal{Z}$ takes a form (6.3). The general case (6.1) follows by induction. By Bayes' rule, the density $\widetilde{\rho}(x_0|\mathcal{Z})\,\mathrm{d}x_0$ is proportional to the product of a prior and the density $\widetilde{\rho}(\mathtt{x}_{\mathtt{obs}}(0+)|x_0, 0+)$ given in eq. (6.8). Completing a square for Gaussian densities then yields:

$$\begin{aligned}
\log\widetilde{\rho}(x_0|\mathcal{Z}) &\propto \log\pi(x_0) + \log\widetilde{\rho}(\mathtt{x}_{\mathtt{obs}}(0+)|x_0, 0+) \\
&\propto -\frac{1}{2}(x_0 - \mu_{\mathtt{pr}})^T \Sigma_{\mathtt{pr}}^{-1}(x_0 - \mu_{\mathtt{pr}}) \\
&\quad -\frac{1}{2}\left(\mathtt{x}_{\mathtt{obs}}(0+) - \widetilde{\mu}_{0+} - \widetilde{L}_{0+}x_0\right)^T \widetilde{M}_{0+}\left(\mathtt{x}_{\mathtt{obs}}(0+) - \widetilde{\mu}_{0+} - \widetilde{L}_{0+}x_0\right) \\
&\propto -\frac{1}{2}\left\{x_0^T\left(\widetilde{L}_{0+}^T\widetilde{M}_{0+}\widetilde{L}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\right)x_0\right. \\
&\qquad \left. -2\left[\left(\mathtt{x}_{\mathtt{obs}}(0+) - \widetilde{\mu}_{0+}\right)^T \widetilde{M}_{0+}\widetilde{L}_{0+} + \mu_{\mathtt{pr}}^T\Sigma_{\mathtt{pr}}^{-1}\right]x_0\right\} \\
&= -\frac{1}{2}\left\{x_0^T\left(\widetilde{H}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\right)x_0 - 2\left(\widetilde{F}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\mu_{\mathtt{pr}}\right)^T x_0\right\} \\
&\propto -\frac{1}{2}\left\{\left[x_0 - \left(\widetilde{H}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\right)^{-1}\left(\widetilde{F}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\mu_{\mathtt{pr}}\right)\right]^T\left(\widetilde{H}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\right)\right. \\
&\qquad \left.\cdot\left[x_0 - \left(\widetilde{H}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\right)^{-1}\left(\widetilde{F}_{0+} + \Sigma_{\mathtt{pr}}^{-1}\mu_{\mathtt{pr}}\right)\right]\right\}.
\end{aligned}$$

□

7

Inference for diffusions from first passage time observations

This chapter is based on joint work with Susanne Ditlevsen and Moritz Schauer.

Bayesian inference methods for diffusion processes are discussed in abundance in the statistics literature within the context of exact observations: $\mathscr{D} := \{x_i \,|\, X_{t_i} = x_i\,;\, i = 1,\ldots,K\}$ (where X denotes the underlying process) as well as partial observations with noise: $\mathscr{D} := \{v_i \,|\, L_i X_{t_i} + \xi_i = v_i,\, i = 1,\ldots,K\}$ (with $L_i \in \mathbb{R}^{d_i \times d}$, $d_i \in \mathbb{N}_+$ and $\xi \sim F_i$ for some distributions F_i, $i = 1,\ldots,K$). However, the topic of inference from first passage time observations is comparatively underrepresented. To the best of my knowledge, to date, only a single Bayesian method that addresses the problem of inference for diffusion processes from their first passage observations has ever been published (Iolov et al., 2017) (not to mention that it is applicable merely to a single—albeit often used—diffusion model: the Ornstein-Uhlenbeck process with time-dependent mean). All of the remaining techniques treating this problem are frequentist and include maximum likelihood estimators obtained via inverting Laplace transforms (Mullowney and Iyengar, 2008) or via numerical approximations to certain representations of the first passage time densities (Zhang et al., 2009), methods of moments (Ditlevsen and Lansky, 2005, 2006) or solutions to Fortet's equations (Ditlevsen and Lansky, 2007; Ditlevsen and Ditlevsen, 2008). Additionally, they are limited to three simple diffusion models: Brownian motion, the Ornstein-Uhlenbeck process and the Cox-Ingersoll-Ross process.

Barring the intractable nature of the first passage time observational setting and the challenge it may present, there is no good reason for its all too seldom presence within the modern statistical discourse, because it arises naturally under a number of experimental designs in biology, survival analysis, physics as well as many other sciences (Ditlevsen and Lansky, 2007; Bachar et al., 2012). However, it is perhaps best known for its relevance to the field of computational neuroscience. Within this field, scientists study time-evolution of neurons' membrane potentials so as to, *inter alia*, understand and decipher communication between those cells (Dayan and

Abbott, 2001). Briefly, membrane potential is the difference in the electric potential between the inside and the outside of a cell and it is generated by the inequality in the concentrations of ions that is inside relative to the outside of a cell. At rest, neurons (depending on their type) are polarised somewhere between —90mV and —40mV (millivolts); however, some neuronal cells are capable of producing very rapid disturbances to this otherwise only moderately fluctuating level, either as a response to an external stimulus or doing so spontaneously (Tuckwell, 1988). These changes manifest themselves in the forms of spikes—periods of a rapid depolarisation, followed by a delayed repolarisation, which, *ceteris paribus*, bring membrane potential back to its resting value (Tuckwell, 1988). An example of the evolution of membrane potential of a neuronal cell is given in fig. 7.1. It illustrates eleven spikes, each lasting around 5ms (milliseconds). In practice, due to well-documented empirical evidence, it is possible to reduce the temporal characterisation of the spiking events to the first passage times of membrane potential to some pre-specified, high enough threshold (Bachar et al., 2012) (a threshold, which, for instance, under the example illustrated by fig. 7.1 could be set to level —20mV). Consequently, an observational setting in which only the times of spike occurrences are recorded are exactly the types of observations I will be considering in this chapter.

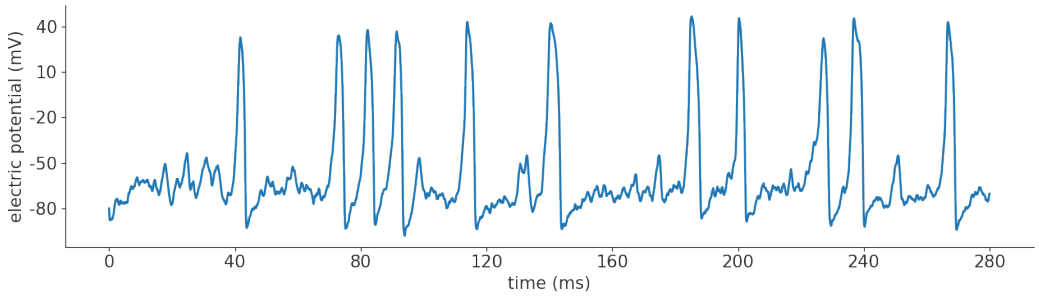


Figure 7.1: Mock-up of the time-evolution of membrane potential, illustrating eleven spikes. Simulated using the stochastic FitzHugh-Nagumo model.

In the subsequent sections I present a comprehensive treatment of the problem of Bayesian inference for diffusion processes from first passage time observations using methods based on data-augmentation. I start from the simplest, one dimensional stochastic *leaky integrate-and-fire* models (Bachar et al., 2012, §5.3), which, presently, are at the limits of the state-of-the-art inference methodologies

(Bachar et al., 2012, §5.5.2). Already at this level the techniques I introduce in this chapter are cable of handling diffusion processes outside of the class discussed in the literature to this date. In appendix A I additionally show how to naturally extend some of the proposed algorithms to multidimensional settings in which first passage times of a single coordinate are observed. Furthermore, it turns out that simulation-based methods are flexible enough to admit extensions to far more complicated models than that—in particular, I show how to tackle certain types of multidimensional, hypoelliptic diffusions, with observations of first passage times of their "smooth" coordinates. These results are of particularly high relevance, because they include the celebrated (stochastic version of) FitzHugh-Nagumo model, devised specifically for modelling neuronal activity (for work applicable to Bayesian inference in a simpler setting of an elliptic version of the FitzHugh-Nagumo model with both coordinates observed at a discrete collection of time-points see for instance Jensen et al. (2012)).

My focus is set primarily on the applications to neuroscience and thus throughout this chapter I lean on the terminology from this field and consider only typical problems relevant to this science; however, it should be made clear that the applicability of the introduced algorithms extends beyond neuroscience and includes some composite observational settings (say, combinations of partial observations and first passage time observations), as well as other diffusion models.

7.1 Diffusion models

7.1.1 Leaky integrate-and-fire

Stochastic leaky integrate-and-fire model, due to its simplicity and despite its limitations, is perhaps the most widely used tool for representing activity of a single neuron (Brunel and Van Rossum, 2007). Because of the mean reversion property and the tractability of the transition densities, the underlying process X is most commonly assumed to follow the dynamics of the Ornstein-Uhlenbeck process or the Cox-Ingersoll-Ross process. Nonetheless, because the novel methodology introduced in this chapter does not share the limitations of the present state-of-the-art

algorithms, I assume in full generality that the underlying process follows the dynamics determined by the stochastic differential equation (1.3)[1], with $d = 1$, until the first stopping time τ (instead of a fixed time T):

$$\begin{aligned} \mathrm{d}X_t = b_\theta(t, X_t)\,\mathrm{d}t + \sigma_\theta(t, X_t)\,\mathrm{d}W_t, \quad X_0 = L_*, \quad t \in [0, \tau], \\ \text{where} \quad \tau := \inf\{t \geq 0 : X_t \geq L^*\}. \end{aligned} \tag{7.1}$$

The Ornstein-Uhlenbeck process and the Cox-Ingersoll-Ross process are just two special cases of the dynamics above. Upon hitting level L^* for the first time, it is assumed that the process takes on a deterministic trajectory that first rapidly depolarises the cell and then returns the process to a reset level L_*. The dynamics are then renewed i.e. X once again follows the dynamics of eq. (7.1) until the first crossing time of level L^* (after renewal). This is illustrated in fig. 7.2. Experimentally, the easiest way to measure the activity of a neuron is to record the times at

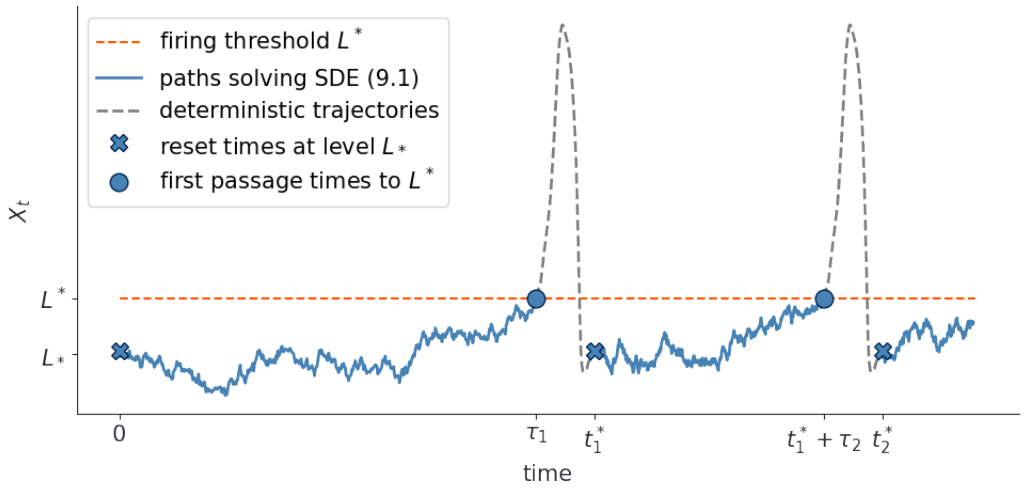


Figure 7.2: Illustration of the leaky integrate-and-fire model. Path X is started from a known level L_* and follows the dynamics from eq. (7.1) until the first time it reaches level L^* (in the plot above this happens at time τ_1). Then, the path takes on a deterministic trajectory that acts out a spike. After known and fixed time ϵ the process is reset at level L_* (in the plot above this happens at time $t_1^* := \tau_1 + \epsilon$). Then, the process once again follows the dynamics from eq. (7.1) until it hits L^* again and the construction is repeated. Assuming that ϵ—the length of the spiking events—is known and only the times of the first passage times are recorded: $\tau_1, t_1^* + \tau_2, \ldots$, without loss of generality one can transform the dataset to $\tau_1, \tau_2, \ldots$, which consists of inter-spike-intervals (ISI) and consider those latter times only.

[1]Because X is defined here as necessarily a one-dimensional process and the methodology remains valid when the coefficients are time-dependent I introduce explicit dependence of the coefficients on t.

which the membrane potential spikes (Tuckwell, 1988) and τ can be taken as a proxy for such spiking events. Then, as shown in fig. 7.2, the data consist of independent and identically distributed realisations of τ—the so-called inter-spike intervals (often abbreviated as ISI):

$$\mathscr{D} := \{\tau_i\,;\, i = 1,\ldots,K\}. \tag{7.2}$$

7.1.2 Multidimensional models

The issue with the leaky integrate-and-fire models is that—by design—the first passage time events are independent and identically distributed, which is contrary to the empirical evidence for the activity of neurons (Dayan and Abbott, 2001). In fact, a number of phenomena contributing towards spiking mechanism are well documented and prompted neuroscientists to posit more complete descriptions of the neuron's behaviour, better captured for instance by the celebrated Hodgkin-Huxley model (Hodgkin and Huxley, 1952). Some of the more complex models exhibit certain important properties—such as direct production of spiking events (without the need for artificial, deterministic trajectories), presence of mechanisms producing clumps of increased spiking activity or extended periods of lack thereof—that are important for working with real data. However, those behaviours cannot be reproduced with simple, one dimensional diffusion models and thus extension to multidimensional processes is necessary.

I assume that the underlying process follows the dynamics given in eq. (1.3)[2], with dimension $d > 1$:

$$\mathrm{d}X_t = b_\theta(X_t)\,\mathrm{d}t + \sigma_\theta(X_t)\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T]. \tag{7.3}$$

This time, the resetting mechanism is acted out by the internal dynamics of the model, without any artificial declarations that the process takes on a deterministic trajectory upon hitting certain pre-specified threshold. In particular, the first passage times are no longer the guarantors of the ensuing spikes; however, it is still reasonable to ascribe them this role, so long as the underlying model in eq. (7.3) is

[2]Now, X can be multidimensional and the usual comment regarding extension to time-inhomogeneous processes applies (see the final paragraphs of section 2.1). For that reason I drop the explicit dependence on time from the notation.

chosen fittingly to the task[3] (Leon et al., 2018). Mathematically, define a series of stopping times:

$$\begin{aligned}\tau_1^* &:= \inf\{t \geq 0 : X_t^{[1]} \geq L^*\},\\ \tau_{*i} &:= \inf\{t \geq \tau_i^* : X_t^{[1]} \leq L_*\}, \quad i = 1,2\ldots,\\ \tau_i^* &:= \inf\{t \geq \tau_{*i-1} : X_t^{[1]} \geq L^*\}, \quad i = 2,3\ldots.\end{aligned} \tag{7.4}$$

τ_{*i}'s, $(i = 1,\ldots)$ denote the renewal times of the first coordinate process $X^{[1]}$ (the times at which diffusion's first coordinate is brought down to the reset level L_* for the first time after spiking). τ_i^*'s, $(i = 1,\ldots)$ denote the first passage times to level L^* of the first coordinate process $X^{[1]}$ since the last renewal times (initially taken to be 0). This is illustrated graphically in fig. 7.3. The observed times of occurrences of spikes are then associated with the stopping times τ_i^*, $(i = 1,\ldots)$. The renewal times τ_{*i}, $(i = 1,\ldots)$ remain latent. Consequently, the data takes the following form:

$$\mathscr{D} := \{\tau_i^*; i = 1,\ldots,K\}. \tag{7.5}$$

Remark 7.1.1. Note that $\tau_i^*; i = 1,\ldots,K$ denote absolute times at which spikes occur. To talk about inter-spike-intervals one should look at $\{\tau_1^*\} \cup \{\tau_{i+1}^* - \tau_i^*, i = 1,\ldots,K-1\}$ instead.

7.1.3 Hypoelliptic models

Hypoelliptic diffusions (i.e. when assumption A3 does not hold, but A4 does) that are treated in this chapter must be of a particular form: they satisfy SDE in eq. (7.3), but the coordinate[4] whose first passage times are observed has a degenerate noise structure. More precisely, (if the first coordinate is the one being observed) the first row of the volatility coefficient in eq. (7.3) describing the dynamics of X is identically equal to a 0-vector, i.e.:

$$\sigma_\theta^{[1,1:d']}(x) = 0_{1\times d'}, \quad \forall x \in \mathbb{R}^d. \tag{7.6}$$

[3]For instance, if a strong oscillatory component is encoded in the diffusion's dynamics, such as is the case with the FitzHugh-Nagumo model.

[4]Or coordinates, as the introduced methodology readily extends to settings in which first passage times of more than one coordinate are observed.

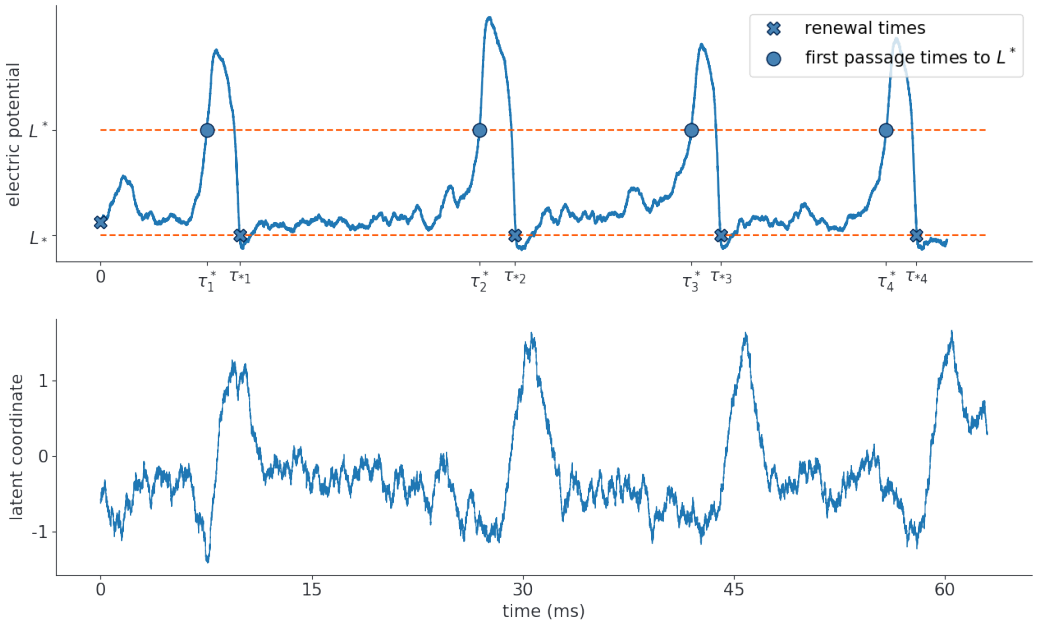


Figure 7.3: Illustration of multidimensional models for neuron's spiking behaviour. Path X is started from some initial position x_0 that remains latent (though, for simplicity, it can be taken to be known) and follows the dynamics from eq. (7.3) throughout the entire time-domain. All coordinates apart from the first one (i.e. the second coordinate in the plot above) are entirely unobserved. Depending on the model, the latent coordinates can have well-defined biological meaning (such as is the case for the Hodgkin–Huxley model, in which the latent coordinates represent various voltage-gated ion channels, leak channels and others (Hodgkin and Huxley, 1952)). Only the times of the first crossing of level L^* by the first coordinate (since the last renewal) are recorded (in the plot above these comprise of τ_i^*, $i = 1,2,3,4$). The renewal times (τ_{*i} $i = 1,2,3,4$ above) remain latent.

Additionally, the drift term of this same coordinate must be a linear function of the process X, i.e. there must exist a vector $\widehat{B} \in \mathbb{R}^d$ and a scalar $\widehat{\beta} \in \mathbb{R}$, so that:

$$b_\theta^{[1]}(x) = \widehat{B}^T x + \widehat{\beta}. \tag{7.7}$$

The former restriction is vital. The latter one might be stronger than necessary, although, it is essential for a rigorous application of the results from this chapter (indeed, in practice, the algorithm presented below works with non-linear drifts as well). See Bierkens et al. (2018) for a conjecture about possible extensions of guided proposals (which is a central component to some of the algorithms of this chapter) to wider class of hypoelliptic diffusions, for which eq. (7.7) need not hold.

The difficulty of treating hypoelliptic diffusions is not restricted to the setting of first passage times observations—the degeneracy of the volatility matrix intro-

duces subtle complications that put such models beyond capabilities of many of the inference algorithms. However, often, the use of such models is backed up by much stronger motivating arguments, than the use of their uniformly elliptic counterparts. For instance, consider a deterministic model based on the following second order, ordinary differential equation, ubiquitous to applications in physics:

$$\begin{cases} \mathrm{d}x_t = \dot{x}_t \,\mathrm{d}t, \\ \mathrm{d}\dot{x}_t = b(x_t, \dot{x}_t)\,\mathrm{d}t, \quad t \in [0, T], \end{cases} \tag{7.8}$$

where $b : \mathbb{R}^2 \to \mathbb{R}$ is some smooth enough function. A physics' interpretation of such ODE is that x_t is a position of a particle and $\dot{x}_t$ is its velocity. The system can be transformed to an SDE by virtue of adding a "noise term" $\sigma\,\mathrm{d}W_t$ to either (or both) of the coordinates. Nonetheless, adding it to the latter coordinate, which results in:

$$\begin{cases} \mathrm{d}X_t = \dot{X}_t \,\mathrm{d}t, \\ \mathrm{d}\dot{X}_t = b(X_t, \dot{X}_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad t \in [0, T], \end{cases} \tag{7.9}$$

conceptually makes more sense (Rogers and Williams, 2000a, Chapter 1, §3.23). Injecting continuous noise directly into the first coordinate would indicate that the state-space that the particle is travelling in experiences some external, continuous, random "shakes", which jitter the particle's position independently of the velocity with which it travels. This is unlikely, as state space is usually understood to be fixed. Instead, any randomness affecting the particle's position comes indirectly, by first affecting its velocity, and then propagating via natural definition of the position as the integrated velocity.

The celebrated FitzHugh-Nagumo model has been designed specifically for applications to neuroscience (FitzHugh, 1961; Nagumo et al., 1962). In the source publication of FitzHugh (1961), it has been defined through a modification to the Van der Pol oscillator (van der Pol and van der Mark, 1928). A natural representation of the latter system takes a form akin to eq. (7.8) and thus it is not surprising that the FitzHugh-Nagumo model can also be transformed to this form (see section 7.4.2). This is why a stochastic version of the FitzHugh-Nagumo model, and by extension, also many other neuronal models, natively take hypoelliptic forms, and why methods capable of dealing with hypoelliptic models are highly relevant to the field of computational neuroscience.

The mathematical description of the spiking behaviour, given in eq. (7.4), as well as the type of collected data, given in eq. (7.5), remain unchanged from section 7.1.2 for hypoelliptic models.

7.2 Simulating diffusions conditioned on first passage times

As I have shown in multiple scenarios in this thesis, simulating appropriately conditioned diffusions is an essential (and also the most challenging) ingredient of the imputation-based, Bayesian inference methods for diffusion processes. In this section (as well as in appendix A), I will describe how to simulate conditioned diffusions corresponding to the observational regimes listed in section 7.1.

7.2.1 Leaky integrate-and-fire models

Consider the simplest setting of leaky integrate-and-fire model in eq. (7.1) and denote with $\mathbb{P}_b^{(\theta)}$ the unconditioned law induced by this SDE. It is easy to see that by the Strong Markov property the conditioning corresponding to the dataset (7.2) is given simply by:

$$\mathcal{Z} := \tau, \tag{7.10}$$

where τ is defined in eq. (7.1), and for the observation set from eq. (7.2) it takes values τ_i, $(i = 1, \ldots, K)$. Simulating diffusion paths from the law $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$ is then done independently on each interval $[0, \tau_i]$, $(i = 1, \ldots, K)$.

7.2.1.1 Rejection sampling via conditioned Wiener law

The first approach that I introduce is an adaptation of the rejection sampler on a path space (see sections 2.3 and 3.1) to the conditioning of the form (7.10). Consequently, assumptions A1, A3, A5–A8 and A10 must hold. Process X is first transformed to the diffusion Y via Lamperti transformation η_θ, which for the time-dependent volatility coefficient is given by:

$$\eta_\theta(t, x) := \int^x \frac{1}{\sigma_\theta(t, u)} \mathrm{d}u.$$

This yields the following stochastic differential equation solved by Y:

$$\begin{aligned} \mathrm{d}Y_t = \alpha_\theta(t, Y_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad Y_0 = L_{*\theta}, \quad t \in [0, \tau(\theta)], \\ \text{where} \quad \tau(\theta) := \inf\{t \geq 0 : Y_t = L^*{}_\theta\}, \end{aligned} \tag{7.11}$$

with $L_{*\theta} := \eta_\theta(0, L_*)$, $L^*{}_\theta := \eta_\theta(\tau, L^*)$ and the drift given by:

$$\alpha_\theta(t,y) := \left[\partial_t \eta_\theta + \frac{b_\theta}{\sigma_\theta} - \frac{1}{2}\partial_y \sigma_\theta\right]\left(\eta_\theta^{-1}(t,y)\right).$$

The problem of sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$ is thus reduced to the problem of sampling from $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$ (with the latter denoting the law induced by eq. (7.11)). As usual, define also $\mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$—the law of Brownian motion conditioned on $\mathcal{Z}$. I will use $\mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$ as the proposal law for $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$. The following theorem then prescribes the likelihood between those two laws.

Theorem 7.2.1. On $\tau < \infty$, the law $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$ is absolutely continuous with respect to $\mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$ and the likelihood between the two is given by:

$$\frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(Y|\mathcal{Z}) = \frac{g_0^{(\theta)}(\tau)}{g_\alpha^{(\theta)}(\tau)} \exp\left\{A_\theta(\tau, L^*{}_\theta) - A_\theta(\tau, L_{*\theta}) - \int_0^\tau \varphi_\theta(s, Y_s)\,\mathrm{d}s\right\},$$

where

$$A_\theta(t,y) := \int^y \alpha_\theta(t,u)\,\mathrm{d}u, \quad \varphi_\theta(t,y) := \left[\alpha_\theta^2 + \partial_y \alpha_\theta + 2\partial_t A_\theta\right](t,y),$$

and where $g_0^{(\theta)}(\tau)$ and $g_\alpha^{(\theta)}(\tau)$ denote the densities of $\mathcal{Z}$ $(= \tau)$ under the proposal Wiener law $\mathbb{P}_0^{(\theta)}$ and the target law $\mathbb{P}_\alpha^{(\theta)}$ respectively:

$$g_0^{(\theta)}(t)\,\mathrm{d}t := \mathbb{P}_0^{(\theta)}(\tau \in \mathrm{d}t), \quad g_\alpha^{(\theta)}(t)\,\mathrm{d}t := \mathbb{P}_\alpha^{(\theta)}(\tau \in \mathrm{d}t).$$

Theorem 7.2.1 paves a clear way for sampling from $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$. As in assumption A10, denote by $l_*(\theta)$ the lower bound on φ_θ and similarly to eq. (2.12), define: $\phi_\theta(t,y) := \varphi_\theta(t,y) - l_*(\theta)$. Then, proposal paths Y are generated from $\mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$ and are accepted with probability $\exp\{-\int_0^\tau \phi_\theta(s, Y_s)\,\mathrm{d}s\}$. Accepted samples are distributed exactly according to $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$ and when transformed via inverse Lamperti transformation η_θ^{-1}, they yield samples from $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$.

There are two ways of executing the accept-reject step. The first solution relies on approximation of $\exp\{-\int_0^\tau \phi_\theta(s, Y_s)\,\mathrm{d}s\}$ via Riemann sums (which prompts for

revealing Y on a dense enough time-grid). The second one is based on the construction involving Poisson point processes, as delineated in section 3.1 (which prompts for revealing Y on a random time-grid). The former solution introduces discretisation errors; however, the latter one is free from them. In view of the algorithms presented in this thesis thus far, the only procedure that requires additional clarifications is the step of sampling from the proposal measure $\mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$.

In order to implement this, notice that $Y \sim \mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$ is simply a Brownian motion conditioned to stay below level $L^*{}_\theta$ on the interval $[0,\tau)$ and conditioned to end up in $L^*{}_\theta$ at the time τ. Define:

$$Z_t := L^*{}_\theta - Y_{\tau-t}, t \in [0,\tau]. \tag{7.12}$$

Then, Z is distributed as standard Brownian motion conditioned to stay positive on $(0,\tau]$ and conditioned to end-up in $L^*{}_\theta - L_{*\theta}$ at time τ. It follows by Williams (1970) and Imhof (1984) that the law of Z coincides with the law of a three dimensional Bessel bridge (joining 0 at time 0 and $(L^*{}_\theta - L_{*\theta})$ at time τ). By Pitman and Yor (1982, 5.d) the path of the latter process admits a decomposition based on three independent Brownian bridges $B := \{B^{[i]}\}_{i=1}^3$ joining 0 and 0 on the time-interval $[0,\tau]$:

$$Z \overset{d}{=} \left\{ \sqrt{\left(B_t^{[1]} + (L^*{}_\theta - L_{*\theta})\frac{t}{\tau}\right)^2 + \left(B_t^{[2]}\right)^2 + \left(B_t^{[3]}\right)^2}, t \in [0,\tau] \right\}. \tag{7.13}$$

Consequently, in order to sample proposal paths $Y \sim \mathbb{P}_0^{(\theta)}(\cdot|\mathcal{Z})$ (in a non-centrally parametrised way), I can first draw three independent, 0–0 Brownian bridges and then transform them via Ψ_θ defined through eqs. (7.12) and (7.13) as follows:

$$\Psi_\theta(B) = \left\{ L^*{}_\theta - \sqrt{\left(B_{\tau-t}^{[1]} + (L^*{}_\theta - L_{*\theta})\frac{\tau-t}{\tau}\right)^2 + \left(B_{\tau-t}^{[2]}\right)^2 + \left(B_{\tau-t}^{[3]}\right)^2}, t \in [0,\tau] \right\}. \tag{7.14}$$

$\Psi_\theta(B)$ may then be used as proposal draws in a rejection sampling setting as candidates for the draws from the target law $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$. This is summarised in algorithm 7.1 below.

Algorithm 7.1 Path space rejection sampler for leaky integrate-and-fire model

while `True` **do**
  for $i = 1, 2, 3$ **do**
    Draw $B^{[i]} \sim \mathbb{W}^*[\tau]$ ▷ 0–0 Brownian bridge on $[0, \tau]$
  Set $Y^\circ \leftarrow \Psi_\theta(B)$ ▷ See eq. (7.14)
  Draw $E \sim \texttt{Exp}(1)$
  if $E > \int_0^\tau \phi_\theta(s, Y_s)\,\mathrm{d}s$ **then** ▷ See sections 2.3 and 3.1 for details
    Set $X \leftarrow \{\eta_\theta^{-1}(s, Y_s), s \in [0, \tau]\}$
    return X ▷ A sample distributed exactly according to $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$

Remark 7.2.1. Notice that neither of $g_0^{(\theta)}(\tau)$ and $g_\alpha^{(\theta)}(\tau)$ is needed for the step of imputing the unobserved path. Nonetheless, the former density will turn out to be essential for the step of updating values of the unknown parameter θ. In preparation for that, denote by $J_a(t)\,\mathrm{d}t$ the first passage time density of a standard Brownian motion to the level a. This density is known in closed form (see for instance Karatzas and Shreve (1998a, §2.8, e.q. 8.5)):

$$J_a(t)\,\mathrm{d}t := \frac{|a|}{\sqrt{2\pi t^3}} \exp\left\{-\frac{a^2}{2t}\right\} \mathrm{d}t. \tag{7.15}$$

Since $g_0^{(\theta)}(\tau)$ is simply the first passage time density to the level $L^*{}_\theta$ of Brownian motion started from $L_{*\theta}$, by translation invariance

$$g_0^{(\theta)}(\tau) = J_{L^*{}_\theta - L_{*\theta}}(\tau). \tag{7.16}$$

7.2.1.2 MCMC on a path space, blocking and other extensions

In the same way that rejection sampling on a path space for diffusions conditioned on an endpoint (described in section 2.3) could be seen as a blueprint for defining a corresponding importance sampler (see section 2.4.1) or an MCMC algorithm on a path space (see algorithm 2.9), the rejection sampler presented in section 7.2.1.1 can act as a starting point for a search of more computationally efficient algorithms. In particular, defining an importance sampler or an independence sampler that is based on algorithm 7.1 proceeds in exactly the same way as it was described in section 2.4.1 and in algorithm 2.9 respectively. For an independence sampler it is further possible to improve the mixing of the chain by, for instance, employing the preconditioned Crank-Nicolson scheme, similarly to how it was done in

section 4.1.6,[5] or by employing blocking, which can be done by following the presentation from section 5.3.[6]

In practice, I found that independence samplers combined with the preconditioned Crank-Nicolson scheme perform exceptionally well across all of the tested settings. I found that in general, first passage time problems that involve scalar diffusions and that are encountered in neuroscience are "simple" enough for the methods based on sampling from conditioned Wiener measures (described in sections 7.2.1.1 and 7.2.1.2) to perform at high efficiency.[7] As I show in section 7.4, because of the very flexible nature of the presented simulation-based methods it is now possible to tackle practical inference problems lying well beyond the capabilities of the state-of-the-art methodologies. This refers not only to applications to a much wider class of diffusion processes, but also to a newly granted flexibility for the experimental design: where combination of various observational settings or integration of multiple experiments that differ in their set-ups can be easily handled by the algorithms. The last point is particularly important for the first passage time observational setting, as the observations from this regime are only weakly informative about the dynamics of the process, much less so than, say, partial observations would be. By pooling the results from multiple experiments that differ in their set-ups and performing joint inference on them it is possible to achieve synergies, increasing the amount of information that each experiment would carry on its own. This last point also extends to other methodologies introduced in this chapter.

[5] More precisely, casting algorithm 7.1 in an MCMC setting and denoting by $B^* := \{B^{*[i]}, i = 1,2,3\}$ the three Brownian bridges accepted at a previous iteration, line 3 of algorithm 7.1 can now be substituted with the two lines below.

1: Draw $B^\circ \sim \mathbb{W}^*[\tau]$ ▷ 0–0 Brownian bridge on $[0,\tau]$
2: Set $B^{[i]} \leftarrow \sqrt{\lambda}B^\circ + \sqrt{1-\lambda}B^{*[i]}$

[6] The version of the non-centrally parametrised rejection sampler with blocking from section 5.3.2 (where blocking is done directly on a non-centrally parametrised space) adapts particularly well to the new setting of section 7.2.1.1: notice that then, each of the three Brownian bridges is simply divided into shorter, overlapping chunks and the paths are drawn consecutively on those segments.

[7] The methods can be tested by the reader on his/her own examples using the code from `https://github.com/mmider/FirstPassageTimeInference.jl`

7.2.2 Multidimensional, uniformly elliptic models

The methods from section 7.2.1 developed for the stochastic leaky integrate-and-fire model can be extended to uniformly elliptic (A3), multidimensional settings. Unfortunately, when applied in practice, the algorithm's performance is underwhelming. The sampler deals with simple, multidimensional diffusions (say linear diffusions from eq. (3.20)), but seems to be unable to efficiently handle highly non-linear cases (say, the uniformly elliptic FitzHugh-Nagumo model). Nonetheless, this outcome should not be surprising: highly non-linear, multivariate diffusions are difficult to deal with even when they are observed exactly on a discrete time-grid. Under conditioning from eq. (7.5), the interval lengths over which diffusion paths need to be imputed are far longer than they are in the vast majority of cases dealt with in the statistics literature today and thus the discussion from section 2.3.4 might provide an intuition for why samplers based on conditioned Wiener measures are not suitable for dealing with highly non-linear diffusion processes conditioned on (7.5).

As a result, due to comparatively low performance of the method crafted for the multidimensional, uniformly elliptic settings and the relatively weaker relevance for the problems at hand, I defer the exposition of the solution for this regime to appendix A. I stress however, that the methodology may still be of interest to applications, which deal with "better behaved" diffusion processes.

7.2.3 Multidimensional, hypoelliptic models

Fortunately, the issues with the uniformly elliptic setting do not impact prospects for solving many problems from neuroscience or physics, where it is the hypoelliptic diffusions that are being studied. Despite inherent difficulties with hypoelliptic diffusions, a different approach to solving the problem yields spectacularly different results, allowing to treat highly non-linear diffusions with great efficiency.

Denote by $\mathbb{P}_b^{(\theta)}$ the unconditioned law induced by eq. (7.3) and notice that the conditioning corresponding to the dataset (7.5) is given by:

$$\mathcal{Z} := \{\tau_i^*;\, i = 1,\ldots,K\}.$$

For simplicity of exposition I assume that $K = 2$, so that:

$$\mathcal{Z} := \{\tau_1^*, \tau_2^*\}, \tag{7.17}$$

the general case follows immediately by induction. Unlike in the simple, scalar model of section 7.2.1, the stopping times τ_i^*, $(i = 1,\ldots,K)$ no longer lead to a factorisation of the target law into independent laws defined on shorter time segments. Instead, global proposals need to be sought.

Hypoellipticity of the form asserted in section 7.1.3, which—due to mathematical complications that arise from it—is normally seen as a burden, awards a crucial simplification for the conditioning given in eq. (7.17). Indeed, it is immediate from eq. (7.6) that the paths of the first coordinate process $X^{[1]}$ are "smoother" than the paths of Brownian motion, because their time-derivatives exist almost everywhere. This is indicative that the conditioning on (7.17) with τ_1^* and τ_2^* taking values τ_1 and τ_2 respectively (with τ_1 and τ_2 some constants) and the conditioning on:

$$\widehat{\mathcal{Z}} := \{X_{\tau_1}^{[1]} = L^*,\ X_{\tau_2}^{[1]} = L^*,\ \exists t \in [\tau_1, \tau_2] \text{ s.t. } X_t \leq L_*\}, \tag{7.18}$$

share a close connection. Under uniform ellipticity, trivially $\mathbb{P}_b(\mathcal{Z}|\widehat{\mathcal{Z}}) = 0$; however, this is no longer the case under hypoellipticity. More precisely, it can be seen from eq. (7.6) that there is positive probability for the first coordinate of a trajectory X conditioned on $\widehat{\mathcal{Z}}$ to have positive time-derivatives in some open balls around τ_1 and τ_2. More rigorously, considering first only a single stopping time, it can be shown that:

Lemma 7.2.1. $\mathbb{P}_b^{(\theta)}(\{\omega \in \Omega : \tau_1^*(\omega) = \tau_1\} | X_{\tau_1}^{[1]} = L^*) > 0$, for all $\tau_1 \in \mathbb{R}_+$, i.e. that when the process X is conditioned to take value L^* at some fixed time τ_1, then the probability that this is the first time that value L^* has been reach is positive.

But then, an extension to arbitrary number of τ_i^*'s follows immediately by Bayes' rule and induction. I summarise this in the following, central result of this section.

Theorem 7.2.2. If (7.6) holds, then $\mathbb{P}_b^{(\theta)}(\mathcal{Z}|\widehat{\mathcal{Z}}) > 0$, where $\mathcal{Z}$ and $\widehat{\mathcal{Z}}$ are as given by eq. (7.17) and eq. (7.18) respectively (and the conditioned-on random variables τ_1^* and τ_2^* in eq. (7.17) take values τ_1 and τ_2 that coincide with the those taken in eq. (7.18)). The same statement remains valid for an arbitrary number of τ_i^*'s.

Theorem 7.2.2 and Bayes' theorem now yield:

$$\frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\widehat{\mathcal{Z}})} = \frac{\mathbb{1}_{\mathcal{Z}}(X)}{\mathbb{P}_b^{(\theta)}(\mathcal{Z}|\widehat{\mathcal{Z}})} \propto \mathbb{1}_{\mathcal{Z}}(X). \tag{7.19}$$

In turn, eq. (7.19) makes it possible to define a suitable rejection sampling algorithm, whose strength lie in reducing the problem of sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$ to a simpler one of sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$. An algorithm for simulating from the latter law has already been proposed in the statistics literature by Bierkens et al. (2018) and it is summarised in section 3.3 of this thesis (see algorithm 3.5 for its extract). The algorithm of Bierkens et al. (2018)—which is based on guided proposals—is well-suited for highly non-linear diffusions of the type encountered in neuroscience. Additionally, I can fully leverage the extensions from chapter 6 to further improve the computational efficiency of the original formulation from Bierkens et al. (2018), as well as use guided proposals in the context of statistical inference for hypoelliptic diffusions—achieving something that has not been explored in Bierkens et al. (2018).

Consequently, modification of Bierkens et al. (2018) to conditioning on the first passage time observations takes a form of an MCMC sampler on a path space, targeting law $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$ and using proposals drawn from the law $\mathbb{P}_{b^\circ}^{(\theta)}$ induced by eq. (3.19), with $\widetilde{h}(t,x) := \mathrm{d}\widetilde{\mathbb{P}}^{(\theta)}(\widehat{\mathcal{Z}}|\widetilde{X}_t = x)$. Notice that the Radon-Nikodým derivative between the target and the proposal laws is given by:

$$\frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X)} = \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\mathcal{Z})}{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\widehat{\mathcal{Z}})} \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\widehat{\mathcal{Z}})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X)} = \frac{\mathbb{1}_{\mathcal{Z}}(X)}{\mathbb{P}_b^{(\theta)}(\mathcal{Z}|\widehat{\mathcal{Z}})} \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X|\widehat{\mathcal{Z}})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X)}. \tag{7.20}$$

Therefore, the Metropolis-Hastings acceptance probability takes a form:

$$\begin{aligned} a_\theta(X^{(n)}, X^\circ) &= 1 \wedge \left[\frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X^\circ|\mathcal{Z})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X^\circ)} \Big/ \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X^{(n)}|\mathcal{Z})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X^{(n)})} \right] \\ &= 1 \wedge \left[\frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X^\circ|\widehat{\mathcal{Z}})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X^\circ)} \mathbb{1}_{\mathcal{Z}}(X^\circ) \Big/ \frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(X^{(n)}|\widehat{\mathcal{Z}})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(X^{(n)})} \right], \end{aligned} \tag{7.21}$$

where I used that $\mathbb{1}_{\mathcal{Z}}(X^{(n)}) \equiv 1$. Equation (7.21) is tractable and its constituent components are given in eq. (3.24), which brings me to the following summary in algorithm 7.2.

Algorithm 7.2 Path imputation for hypoelliptic diffusions conditioned on FPTs

`accepted` $\leftarrow$ `False`
repeat ▷ Initialisation
  Draw $X^{(0)} \sim \mathrm{d}\mathbb{P}^{(\theta)}_{b^\circ}$ ▷ As per eq. (3.19), with $\widetilde{h}(t,x) := \mathrm{d}\widetilde{\mathbb{P}}^{(\theta)}(\widehat{\mathcal{Z}}|\widetilde{X}_t = x)$
  if $\mathcal{Z}$ occurs **then**
    `accepted` $\leftarrow$ `True`
until `accepted`
for $n = 1,\ldots,N$ **do** ▷ MCMC on a path space
  Draw $X^\circ \sim \mathrm{d}\mathbb{P}^{(\theta)}_{b^\circ}$ ▷ As per eq. (3.19), with $\widetilde{h}(t,x) := \mathrm{d}\widetilde{\mathbb{P}}^{(\theta)}(\widehat{\mathcal{Z}}|\widetilde{X}_t = x)$
  Draw $U \sim$ `Unif([0,1])`
  if $U \leq a_\theta(X^{(n-1)}, X^\circ)$ **then** ▷ See eq. (7.21)
    Set $X^{(n)} \leftarrow X^\circ$
  else
    Set $X^{(n)} \leftarrow X^{(n-1)}$
return $\{X^{(n)}; n = 1,\ldots,N\}$ ▷ Markov chain with invariant law $\mathbb{P}^{(\theta)}_b(\cdot|\mathcal{Z})$

Naturally, in view of the central application to Bayesian inference, a non-centrally parametrised version of algorithm 7.2 is needed. In example 4.1.2 and in algorithm 6.2 I have already described how to define a non-centrally parametrised version of guided proposals for conditioning of the form (6.1). Amending algorithm 7.2 so as to introduce the main ideas from algorithm 6.2 follows trivially. Afterwards, optional implementation of the preconditioned Crank-Nicolson scheme and/or blocking follows as per descriptions in section 4.1.6 and section 5.3 respectively.

7.3 Inference

The samplers of conditioned diffusions introduced in the previous section make it possible to define Bayesian inference algorithms based on data augmentation for the models listed in section 7.1. I assume that $\theta \in \Theta$ is random and comes equipped with a prior density $\pi(\theta)$. The aim is to derive the posterior density:

$$\pi(\theta|\mathscr{D}) \propto \pi(\mathscr{D}|\theta)\pi(\theta).$$

The approach I take here follows the one delineated in chapter 4. I define a Gibbs sampler, which alternately imputes the unobserved parts of path (by sampling from $\pi(X|\theta,\mathscr{D})$) and updates unknown parameter θ (by sampling from $\pi(\theta|X,\mathscr{D})$). The invariant density of the resulting chain $\{(\theta^{(n)}, X^{(n)}); n = 1,\ldots,N\}$ is $\pi(\theta, X|\mathscr{D})$, whereas in the case of the marginal chain $\{\theta^{(n)}; n = 1,\ldots,N\}$ it is $\pi(\theta|\mathscr{D})$.

7.3.1 Leaky integrate-and-fire models

In this case, the imputation density $\pi(X|\theta,\mathscr{D})$ is given by $\bigotimes_{i=1}^{K} \mathrm{d}\mathbb{P}_{b}^{(\theta)}(\cdot|\mathcal{Z}_i)$, with $\mathcal{Z}_i$, $(i=1,\ldots,K)$ defined in eq. (7.10) for each observation τ_i, $(i=1,\ldots,K)$. As discussed in section 7.2.1, sampling from $\mathbb{P}_{b}^{(\theta)}(\cdot|\mathcal{Z}_i)$ on each sub-interval $[0,\tau_i]$, $(i=1,\ldots,K)$ can be completed by either of the following:

- Exact rejection sampling on a path space (as per algorithm 7.1), with Poisson point process construction completing the accept-reject step (as per section 3.1)
- Rejection sampling on a path space (as per algorithm 7.1), with left-Riemann sum approximations to the integrals in the accept-reject step
- Independence sampler using proposal draws from $\mathbb{P}_{0}^{(\theta)}(\cdot|\mathcal{Z}_i)$ to target $\mathbb{P}_{\alpha}^{(\theta)}(\cdot|\mathcal{Z}_i)$
- The Metropolis-Hastings step using blocked proposals from $\mathbb{P}_{0}^{(\theta)}(\cdot|\mathcal{Z}_i)$ to target $\mathbb{P}_{\alpha}^{(\theta)}(\cdot|\mathcal{Z}_i)$—updates of each block can be completed with either of the techniques from the former three bullet points

Additionally, the latter two algorithms may be modified with the preconditioned Crank-Nicolson scheme.

7.3.1.1 Unbiased inference

If an "exact" version of the sampler is chosen (this refers to the first bullet point from the list above or the last one, so long as each block is updated by the exact rejection sampler on a path space), then it is possible to preserve this exactness and define an inference algorithm devoid of any discretisation errors, just as it was done in section 4.2. To this end, assume for simplicity that samples from $\mathbb{P}_{b}^{(\theta)}(\cdot|\mathcal{Z}_i)$ are drawn according to the exact rejection sampler of Beskos et al. (2008) (see section 3.1 for details) that is appropriately modified for algorithm 7.1 (a version of the algorithm with blocking follows analogously). Then, each accepted imputation sample $X := \{X^{[i]}\}_{i=1}^{K}$ (which consist of K independent trajectories $X^{[i]}$ defined on the time-intervals $[0,\tau_i]$, $i=1,\ldots,K$, that I collect into a single object X for brevity) has been revealed at a finite collection of random time points

$$\{X^{[i]}_{\chi_j^{[i]}}; j=1,\ldots,\varkappa_i\}, \quad (i=1,\ldots,K), \tag{7.22}$$

coinciding with the times of a simulated Poisson point process $\Phi := \{\Phi_i,\ i = 1,\ldots,K\}$. To be more precise, since a non-centrally parametrised version of the algorithm has to be used, it is in fact the three Brownian bridges from line 3 of algorithm 7.1 that are being revealed at times $\chi_j^{[i]}$, $(j = 1,\ldots,\varkappa)$ on each time-interval $[0,\tau_i]$, $(i = 1,\ldots,K)$:

$$\{B^{[k,i]}_{\chi_j^{[i]}}; k = 1,2,3;\ j = 1,\ldots,\varkappa_i\}, \quad (i = 1,\ldots,K), \tag{7.23}$$

and path X in eq. (7.22) is then derived as a by-product via transformation $\Psi_\theta^{[i]}$ defined for each interval $[0,\tau_i]$, $(i = 1,\ldots,K)$ analogously to eq. (7.14).

Additionally, other random variables Υ_i, $(i = 1,\ldots,K)$ (such as layered information) might have been simulated along the way. For notational convenience denote with ν_i, $i = 1,\ldots,K$ the absolute times of the renewal times associated with the beginnings of the intervals $[0,\tau_i]$, $i = 1,\ldots,K$ (i.e. so that the absolute times of spike occurrences are given by $\nu_i + \tau_i$ $(i = 1,\ldots,K)$) and write $L_{*\theta}^{[i]} := \eta_\theta(\nu_i, L_*)$ and $L_\theta^{*[i]} := \eta_\theta(\nu_i + \tau_i, L^*)$, $(i = 1,\ldots,K)$. Then, similarly to Sermaidis et al. (2013, theorem 1), I can show the following.

Theorem 7.3.1. Under a leaky integrate-and-fire model and an assumption of sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z}_i)$, $(i = 1,\ldots,K)$ being completed via exact rejection sampling on a path space, the joint density for the data $\mathscr{D}$, parameter θ and simulated variables

$$\mathscr{S} := \{\{B^{[k,i]}_{\chi_j^{[i]}}; k = 1,2,3;\ j = 1,\ldots,\varkappa_i\} \cup \Upsilon_i\}_{i=1}^K, \tag{7.24}$$

is given by

$$\begin{aligned}\pi(\mathscr{S},\theta,\mathscr{D}) = \pi(\theta)\prod_{i=1}^K \Bigg\{ &\exp\left\{A_\theta(\nu_i+\tau_i, L_\theta^{*[i]}) - A_\theta(\nu_i, L_{*\theta}^{[i]})\right\} J_{L_\theta^{*[i]} - L_{*\theta}^{[i]}}(\tau_i) \\ &\cdot [l_\theta^*(\Upsilon_i)]^{\varkappa_i} \cdot \exp\{-[l_{*\theta} - 1 + l_\theta^*(\Upsilon_i)]\tau_i\}\, \mathbb{1}_{\mathcal{Z}_i}\left(\Psi_\theta^{[i]}(B^{[1:3,i]})\right) \\ &\cdot \prod_{j=1}^{\varkappa_i}\left[1 - \phi_\theta(\nu_i + \chi_j^{[i]}, \Psi_\theta^{[i]}(B^{[1:3,i]})_{\chi_j^{[i]}})/l_\theta^*(\Upsilon_i)\right]\Bigg\},\end{aligned} \tag{7.25}$$

where l_θ^* and $l_{*\theta}$ are defined in assumptions A10 and A12 respectively and where J is given by eq. (7.15).

With theorem 7.3.1 in place, it possible to update parameter θ by employing the Metropolis-Hastings algorithm: proposing a new value for the parameter θ from some transition kernel $q(\theta^{(n)}, \theta^\circ)$ and accepting it with the probability

$$a(\theta^{(n)}, \theta^\circ) = 1 \wedge \left[\frac{\pi(\mathscr{S}^{(n)}, \theta^\circ, \mathscr{D}) q(\theta^\circ, \theta^{(n)})}{\pi(\mathscr{S}^{(n)}, \theta^{(n)}, \mathscr{D}) q(\theta^{(n)}, \theta^\circ)} \right].$$

The expression for the acceptance probability $a(\theta^{(n)}, \theta^\circ)$ involves only a finite number of variables, all of which are known and thus it can be evaluated without making any approximation errors. Naturally, the commentary applicable to the algorithm of Sermaidis et al. (2013), summarised in the second part of section 4.2 still applies here, with perhaps the most important point being that for a new sample θ° it is possible that $l^*_{\theta^\circ}(\Upsilon_i) > l^*_\theta(\Upsilon_i)$. This means that, potentially, an extra Poisson point process on $[0, \tau_i] \times [l^*_\theta(\Upsilon_i), l^*_{\theta^\circ}(\Upsilon_i)]$ needs to be sampled, followed by revealing an already accepted path X at additional time points. This is a suboptimal strategy and for more details on this issue and a more practical way of revealing path X, see algorithm 4.9 in section 4.2 or Sermaidis et al. (2013, §4.1). The unbiased inference algorithm for stochastic leaky integrate-and-fire models with first passage time observations is summarised in algorithm 7.3 below.

Algorithm 7.3 *Exact* inference for stochastic leaky integrate-and-fire models from first passage times (FPTs)

1: Initialise $\theta^{(0)}$
2: **for** $n = 1, \ldots, N$ **do**
3: Draw $X \sim \bigotimes_{i=1}^{K} \mathbb{P}_b^{(\theta^{(n-1)})}(\cdot | \mathcal{Z}_i)$, effectively revealing $\mathscr{S}$ in eq. (7.24) ▷ Use non-centrally parametrised, exact version of algorithm 7.1
4: Draw $\theta^\circ \sim q(\theta, \cdot)$
5: Draw $U \sim$ `Unif([0,1])`
6: **if** $U \leq \frac{\pi(\theta^\circ, \mathscr{S}, \mathscr{D}) q(\theta^{(n-1)}, \theta^\circ)}{\pi(\theta^{(n-1)}, \mathscr{S}, \mathscr{D}) q(\theta^\circ, \theta^{(n-1)})}$ **then** ▷ see eq. (7.25)
7: Set $\theta^{(n)} \leftarrow \theta^\circ$
8: **else**
9: Set $\theta^{(n)} \leftarrow \theta^{(n-1)}$
10: **return** $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with the invariant density $\pi(\theta | \mathscr{D})$

7.3.1.2 Inference with an approximation step

As an alternative to algorithm 7.3, an approximate procedure can be employed, in which imputation via algorithm 7.1 (or via an independence sampler or the

Metropolis-Hastings algorithm derived from algorithm 7.1) employs approximations to the integrals via left-Riemann sums. Then, the following counterpart to theorem 7.3.1 gives a way to update the parameter θ:

Theorem 7.3.2. Under leaky integrate-and-fire model, the joint density under the target measure for the data $\mathscr{D}$, parameter θ and the non-centrally parametrised imputation variables $B := \{B^{[k,i]},\ k = 1,2,3,\ i = 1,\ldots,K\}$ is given by:

$$\pi(\theta,B,\mathscr{D}) = \pi(\theta)\prod_{i=1}^{K}\Bigg\{ \exp\Big\{A_\theta(\nu_i+\tau_i,L^{*[i]}_\theta) - A_\theta(\nu_i,L^{[i]}_{*\theta})\Big\} J_{L^{*[i]}_\theta - L^{[i]}_{*\theta}}(\tau_i) \\ \cdot \exp\left\{-\int_0^{\tau_i}\varphi_\theta\Big[\nu_i+t,\Psi^{[i]}_\theta(B^{[1:3,i]})_t\Big]\,\mathrm{d}t\right\}\mathbb{1}_{\mathcal{Z}_i}\Big(\Psi^{[i]}_\theta(B^{[1:3,i]})\Big)\Bigg\} \tag{7.26}$$

Evaluation of the joint density from theorem 7.3.2 can once again be completed with approximations of the integrals via left-Riemann sums and thus the updates of the parameter θ can be completed with the Metropolis-Hastings algorithm. The scheme is summarised in algorithm 7.4 below. For simplicity I assume that the path is updated via independence sampler using proposals from $\mathbb{P}^{(\theta)}_0(\cdot|\mathcal{Z}_i)$ to target $\mathbb{P}^{(\theta)}_\alpha(\cdot|\mathcal{Z}_i)$, although in full generality any other scheme from the bullet-point list above can be employed.

7.3.2 Multidimensional, hypoelliptic models

For the reasons listed in section 7.2.2 I defer the exposition of an inference algorithm suitable for multidimensional, uniformly elliptic models until appendix A and instead focus directly on the hypoelliptic setting. The general outline of the inference algorithm remains unchanged from the previous section—only its individual steps need to be adjusted so as to accommodate the results from section 7.2.3.

Naturally, the imputation step is performed according to a non-centred version of algorithm 7.2. It is the parameter update step that needs further explanations. First, recall that under this observational regime $\mathcal{Z}$ and $\widehat{\mathcal{Z}}$ are respectively given by:

$$\mathcal{Z} := \{\tau^*_i;\ i = 1,\ldots,K\},\ \text{where } \tau^*_i\text{'s are stopping times defined in eq. (7.4)},$$
$$\widehat{\mathcal{Z}} := \{X^{[1]}_\tau = L^*, \tau \in \{\tau^*_1,\ldots,\tau^*_K\} \text{ and } \forall i = 1,\ldots,K\ \exists t \in [\tau^*_{i-1},\tau^*_i] \text{ s.t. } X_t \le L_*\},$$
$$\text{where } \tau^*_i\text{'s are simply constants, and } \tau^*_0 := 0.$$

Algorithm 7.4 Inference for stochastic leaky integrate-and-fire models from FPTs

Initialise $\theta^{(0)}$
Draw $Y^{(0)} \sim \bigotimes_{i=1}^{K} \mathbb{P}_0^{(\theta^{(0)})}(\cdot|\mathcal{Z}_i)$ (revealing $B^{(0)}$ on a dense enough time grid)
for $n = 1, \ldots, N$ **do**
    for $i = 1, \ldots, K$ **do**
        for $k = 1, 2, 3$ **do**
            Draw $B^{\circ[k,i]} \sim \mathbb{W}^*[\tau_i]$ ▷ 0–0 Brownian bridge on $[0, \tau_i]$
        Set $Y^\circ \leftarrow \Psi^{[i]}_{\theta^{(n-1)}}(B^{\circ[1:3]})$
        Draw $E \sim \texttt{Exp(1)}$
        if $E \geq \int_0^{\tau_i} \phi_{\theta^{(n-1)}}(\nu_i + s, Y_s^\circ)\,\mathrm{d}s$ **then**
            Set $B^{(n)[1:3,i]} \leftarrow B^{\circ[1:3,i]}$
        else
            Set $B^{(n)[1:3,i]} \leftarrow B^{(n-1)[1:3,i]}$
    Draw $\theta^\circ \sim q(\theta^{(n-1)}, \cdot)$
    Draw $U \sim \texttt{Unif([0,1])}$
    if $U \leq \frac{\pi(\theta^\circ, B^{(n)}, \mathscr{D}) q(\theta^{(n-1)}, \theta^\circ)}{\pi(\theta^{(n-1)}, B^{(n)}, \mathscr{D}) q(\theta^\circ, \theta^{(n-1)})}$ **then** ▷ See eq. (7.26)
        Set $\theta^{(n)} \leftarrow \theta^\circ$
    else
        Set $\theta^{(n)} \leftarrow \theta^{(n-1)}$
return $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with the invariant density $\pi(\theta|\mathscr{D})$

The following result is instrumental:

Theorem 7.3.3. For the hypoelliptic models of the form asserted in section 7.1.3, the joint density for the observations $\mathscr{D}$, the parameter θ and the non-centrally parametrised imputation variables W is given by:

$$\pi(\theta, W, \mathscr{D}) = \pi(\theta)\widetilde{h}_\theta(0, x_0) \exp\left\{\int_0^{L^*_K} G_\theta(\Psi_\theta(W)_t)\,\mathrm{d}t\right\} \mathbb{1}_{\mathscr{D}}(\Psi_\theta(W)),$$

where $\widetilde{h}_\theta(0, x_0) := \mathrm{d}\widetilde{\mathbb{P}}^{(\theta)}(\widehat{\mathcal{Z}}|X_t = x)$, and where G_θ is defined in proposition 3.3.1.

Notice that $\mathrm{d}\widetilde{\mathbb{P}}^{(\theta)}(\widehat{\mathcal{Z}}|X_t = x)$ above denotes the density that concerns partial observations and not first passage time observations. This means that the results from chapter 6 can be used to compute this term. Consequently, $\pi(\theta, W, \mathscr{D})$ above is tractable and to update θ it is possible to employ the Metropolis-Hastings algorithm analogously to how it was done in section 7.3.1, except using the density from theorem 7.3.3 whenever appropriate. The full procedure is summarised by algorithm 7.5.

Algorithm 7.5 Inference from first passage times for hypoelliptic diffusions

Initialise $\theta^{(0)}$
`accepted` $\leftarrow$ `true`
repeat ▷ Path initialisation
  Draw $W^{(0)} \sim \bigotimes_{i=1}^{K} \mathbb{W}^*([\tau_{i-1}^*, \tau_i^*])$ ▷ indep. Brownian motions on $[\tau_{i-1}^*, \tau_i^*]$
  Set $X^{(0)} \leftarrow \Psi_{\theta^{(0)}}(W^{(0)})$
  if $\mathcal{Z}$ occurs **then**
    `accepted` $\leftarrow$ `True`
until `accepted`
for $n = 1, \ldots, N$ **do** ▷ MCMC algorithm
  Draw $W^\circ \sim \bigotimes_{i=1}^{K} \mathbb{W}^*([\tau_{i-1}^*, \tau_i^*])$ ▷ indep. Brownian motions on $[\tau_{i-1}^*, \tau_i^*]$
  Set $X^\circ \leftarrow \Psi_{\theta^{(n-1)}}(W^\circ)$
  Draw $U \sim$ `Unif([0,1])`
  if $U \leq a_{\theta^{(n-1)}}(X^{(n-1)}, X^\circ)$ **then** ▷ a is defined in eq. (7.21)
    Set $W^{(n)} \leftarrow W^\circ$
  else
    Set $W^{(n)} \leftarrow W^{(n-1)}$
  Draw $\theta^\circ \sim q(\theta^{(n-1)}, \cdot)$
  Draw $U \sim$ `Unif([0,1])`
  if $U \leq \frac{\pi(\theta^\circ, W^{(n)}, \mathscr{D}) q(\theta^{(n-1)}, \theta^\circ)}{\pi(\theta^{(n-1)}, W^{(n)}, \mathscr{D}) q(\theta^\circ, \theta^{(n-1)})}$ **then** ▷ See the density in theorem 7.3.3
    Set $(\theta^{(n)}, X^{(n)}) \leftarrow (\theta^\circ, \Psi_{\theta^\circ}(W^{(n)}))$
  else
    Set $(\theta^{(n)}, X^{(n)}) \leftarrow (\theta^{(n-1)}, \Psi_{\theta^{(n-1)}}(W^{(n)}))$
return $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with the invariant density $\pi(\theta|\mathscr{D})$

7.4 Numerical results

I start this section with the treatment of stochastic, leaky integrate-and-fire models. I show that the problem of inference from first passage time observations of the Ornstein-Uhlenbeck process can be easily handled with the introduced methodology. First passage time observations are not informative enough to make it possible to conduct joint inference on both parameters in the drift (an issue that is recognised in the literature); however, any one of the two drift parameters can be estimated jointly with the volatility coefficient. I then consider a modified problem, in which the underlying process is assumed to follow the dynamics of the Ornstein-Uhlenbeck process that is modified by inclusion of an extra, time-dependent term in its drift. This term could be interpreted as the external modulation of the synap-

tic input exerted by the scientist (which he or she has a direct control over) with the aim of changing the voltage evolution of a cell. Already in this setting simultaneous inference of all parameters is possible; however, for the didactic purposes I consider a non-standard set-up (which, arguably, might be more relevant to fields other than neuroscience) in which data is pooled from three independent experiments with different levels of firing and renewal thresholds and show that all three parameters can be estimated jointly. Finally, I consider a non-standard diffusion model—a Langevin diffusion with t-distribution as its invariant measure—to point to the fact that the introduced methodology extends to an array of other diffusion models that might have not been considered previously due to their intractability.

In the second part of this section I consider hypoelliptic diffusions and focus on an application to the FitzHugh-Nagumo model. To the best of my knowledge, despite the relevance of this system to neuroscience, the problem of inference for it from first passage time observations has not been treated in the literature.

7.4.1 Leaky integrate-and-fire models

7.4.1.1 The Ornstein-Uhlenbeck process

I assume that the underlying process X from eq. (7.2) is given by the Orsnetin-Uhlenbeck process. Equation (7.2) then becomes

$$\begin{aligned} \mathrm{d}X_t &= \theta^{[1]}(\theta^{[2]} - X_t)\,\mathrm{d}t + \theta^{[3]}\,\mathrm{d}W_t, \quad X_0 = L_*, \quad t \in [0,\tau], \\ &\text{where} \quad \tau := \inf\{t \geq 0 : X_t \geq L^*\}. \end{aligned} \tag{7.27}$$

I set $(L_*, L^*) = (0, 10)$ and $\theta = (0.1, 15, 1)^T$ and simulated $K = 30$ observations of first passage times; they are plotted in fig. 7.4. For this example, the dura-

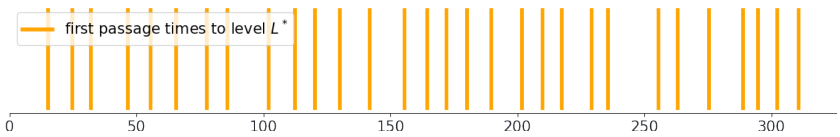


Figure 7.4: First passage time observations of the Ornstein-Uhlenbeck process

tion of spikes, as well as the absolute times at which the intervals start are irrelevant and only the duration of the inter-spike-intervals matter. Therefore, without loss of generality I assume that at the moment of the first passage time to level

L^* the membrane potential instantaneously jumps to renewal level L_* and starts anew with the dynamics from eq. (7.27), see fig. 7.5 for an illustration. Simulta-

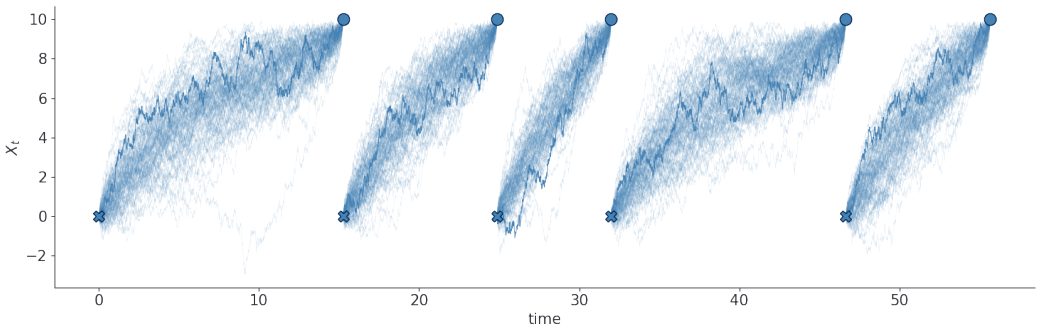


Figure 7.5: Imputed path of the Ornstein-Uhlenbeck process under first-passage time observational regime. Samples of 100 paths on the first 5 inter-spike-intervals, imputed during inference for $\theta^{[1]}$ and $\theta^{[3]}$. The spikes were removed due to irrelevance for this model.

neous estimation of the two parameters present in the drift term—$\theta^{[1]}$ and $\theta^{[2]}$—appears to be impossible due to identifiability issues. This observation is consistent with the remarks made in the statistics literature (Ditlevsen and Lansky, 2007; Iolov et al., 2017). I applied algorithm 7.4 to conduct two separate inference procedures: the first one has been done for parameters $(\theta^{[1]}, \theta^{[3]})$ (assuming $\theta^{[2]}$ to be known), the second one for $(\theta^{[2]}, \theta^{[3]})$ (assuming $\theta^{[1]}$ to be known). I used improper priors $\pi(\theta^{[1]}, \theta^{[3]}) \propto \frac{1}{\theta^{[1]}\theta^{[3]}}$ in the former case and $\pi(\theta^{[2]}, \theta^{[3]}) \propto \frac{1}{\theta^{[3]}}$ in the latter. For updates of $\theta^{[2]}$ I used random-walk proposals $q(x, \cdot) = x + U$, with $U \sim \texttt{Unif}([-\epsilon, \epsilon])$. For the other two parameters I used an exponentiated random-walk: $q(x, \cdot) \sim xe^U$, $U \sim \texttt{Unif}([-\epsilon, \epsilon])$. ϵ parameter was tuned for each parameter separately to target acceptance rate ~ 0.23. Neither of the preconditioned Crank-Nicolson scheme nor blocking were used, as the imputation acceptance rate stayed above 80% across most inter-spike-intervals. Figure 7.5 shows a sample of paths that were imputed by the inference algorithm on the first 5 inter-spike-intervals. Figure 7.6 gives the results of inference for parameters $\theta^{[1]}$ and $\theta^{[3]}$. The results of inference for parameters $\theta^{[2]}$ and $\theta^{[3]}$ are given in fig. 7.7.

7.4.1.2 A modified Ornstein-Uhlenbeck process

Let the underlying process X from eq. (7.2) be given by the Orsnetin-Uhlenbeck process that is modified by a presence of an extra, time-dependent term. Equa-

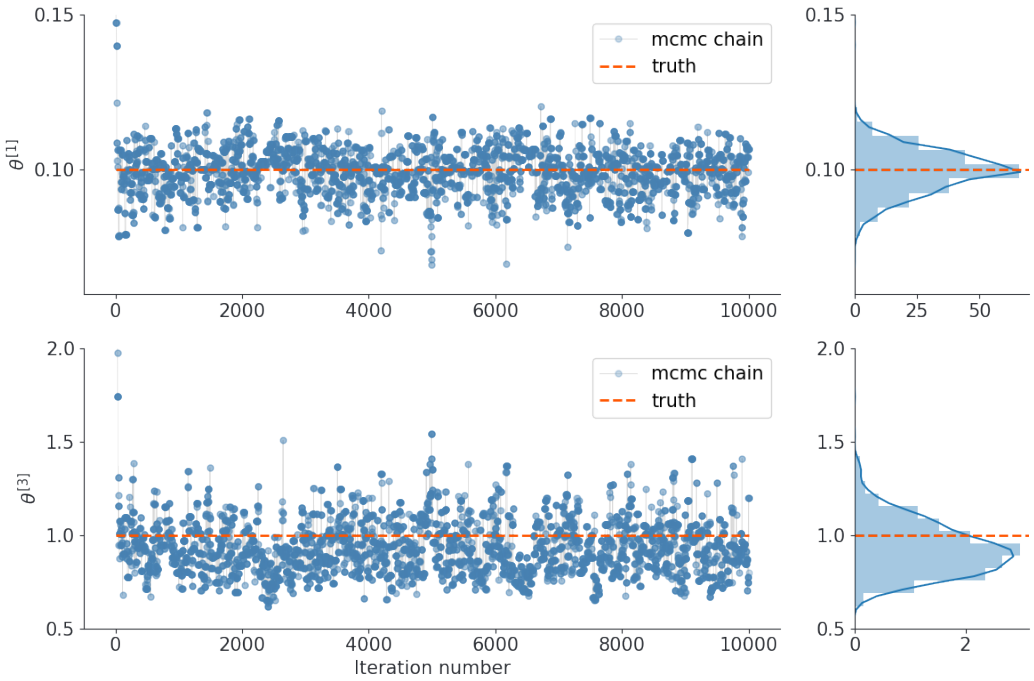


Figure 7.6: Traceplots for the inference on the Ornstein-Uhlenbeck process from first passage time observations. The chain of $\theta^{[1]}$ and $\theta^{[3]}$ was initialised at $\theta^{[1](0)}, \theta^{[3](0)} = (0.3, 3)$.

tion (7.2) takes the following form

$$\mathrm{d}X_t = (\theta^{[1]} + f(\nu_i + t) - \theta^{[2]} X_t)\mathrm{d}t + \theta^{[3]}\mathrm{d}W_t, \quad X_0 = L_*, \quad t \in [0, \tau],$$
$$\text{where} \quad \tau := \inf\{t \geq 0 : X_t \geq L^*\}.$$

For the experiments I took $f(t) = 0.5 \sin\left(\frac{\pi t}{20}\right)$ (note that in the above SDE, $f(\nu_i + t)$ is started from ν_i—the absolute time of the renewal that initiated given inter-spike-interval). Notice that I used different parametrisation of the Ornstein-Uhlenbeck process from the one presented in section 7.4.1.1—this is for the reasons related to conjugate updates that I address in more detail below. I set $\theta = (0.1, 1.5, 1)^T$ and performed three independent experiments with $({L_*}^{[1]}, {L^*}^{[1]}) = (0, 10)$, $({L_*}^{[2]}, {L^*}^{[2]}) = (0, 5)$ and $({L_*}^{[3]}, {L^*}^{[3]}) = (5, 8)$ respectively, each time simulating 40 first passage time observations. I set the length of each spiking event to 1. The observations from all three experiments are plotted in fig. 7.8. For inference I set improper prior on the parameter $\theta^{[3]}$: $\pi(\theta^{[3]}) \propto \frac{1}{\theta^{[3]}}$ and independent Gaussian priors on $\theta^{[1]}$ and $\theta^{[2]}$, both with mean 0 and variance 1000. The updates of $\theta^{[3]}$ were completed with an exponentiated random-walk, as in section 7.4.1.1; however, for the joint updates

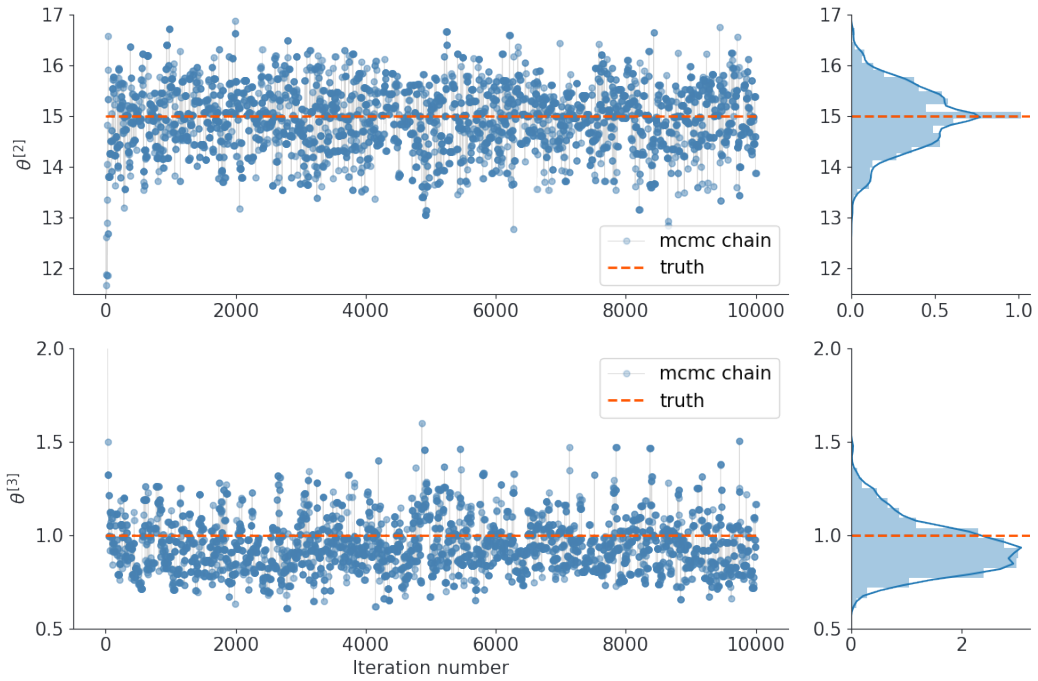


Figure 7.7: Traceplots for the inference on the Ornstein-Uhlenbeck process from first passage time observations. The chain of $\theta^{[2]}$ and $\theta^{[3]}$ was initialised at $\theta^{[2](0)}, \theta^{[3](0)} = (10, 3)$.

of $\theta^{[1:2]}$ I employed conjugate updates whose form is easily derivable from example 4.1.1 by introducing modifications that accommodate presence of the function $f(t)$. It is possible to employ the Metropolis-within-Gibbs algorithm for the updates of $\theta^{[1]}$ and $\theta^{[2]}$, but due to severe correlation between the two parameters, the mixing of the resulting chain is a couple of orders of magnitudes slower than the one based on the conjugate updates. The results of inference are given in fig. 7.9. All three parameters have been successfully identified.

If the inference algorithm were to be re-run with a sufficiently large number of observations from the three experiments, but for a model without a time-dependent term (i.e. using the classical Ornstein-Uhlenbeck process in eq. (7.27)), then identification of all three parameters would have still been possible. Vice-versa, if only one experiment with sufficiently large number of observations were conducted (so that L^* and L_* were fixed for all observations), then identification of all three parameters would have been possible as well. The combination of various firing thresholds and inclusion of the time-dependent term were chosen to illustrate the flexibility of the introduced methodology.

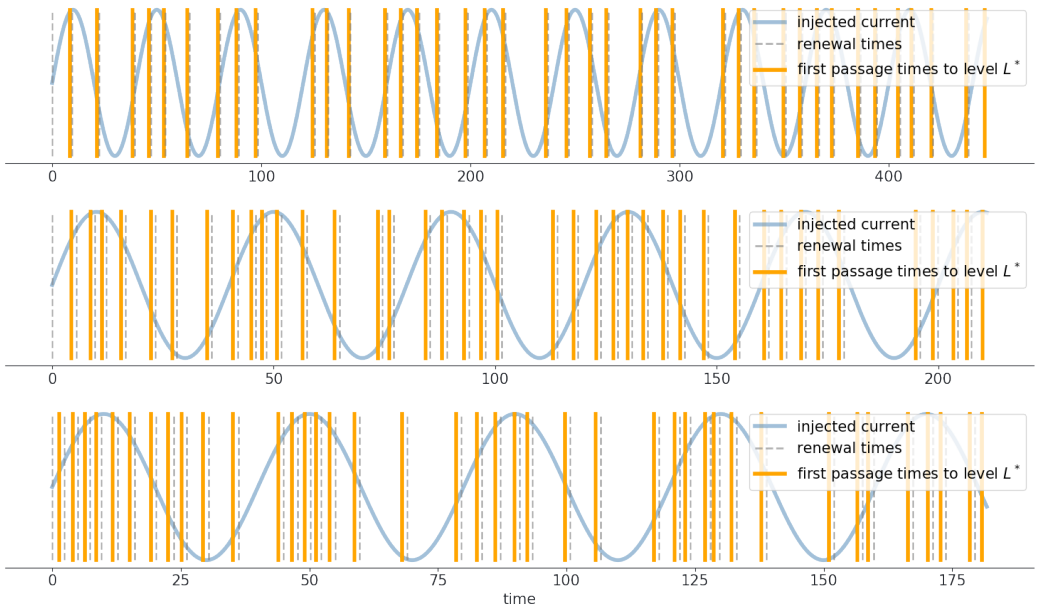


Figure 7.8: First passage time observations of the modified Ornstein-Uhlenbeck process from three independent experiments. $({L_*}^{[1]}, {L^*}^{[1]}) = (0, 10)$ was used to produce the top plot, $({L_*}^{[2]}, {L^*}^{[2]}) = (0, 5)$ for the middle and $({L_*}^{[3]}, {L^*}^{[3]}) = (5, 8)$ for the bottom. Curves proportional to the injected current as a function of time are plotted in the background. Notice the clumping behaviour that arises as a by-product of the correlation between the injected current and the spike production.

Furthermore, it takes only a slight change of perspective to interpret the presence of a time-dependent term in the drift in an entirely different way. Notice that if $X^{(f)}$ denotes the diffusion from eq. (7.27), $X^{(0)}$ the corresponding diffusion with $f(t) \equiv 0$ and $F(t) := \int_0^t f(s)\,\mathrm{d}s$, then $\mathrm{d}(X_t^{(f)} - F(t)) = \mathrm{d}X_t^{(f)} - f(t)\,\mathrm{d}t = \mathrm{d}X_t^{(0)}$ and therefore, since $F(0) = 0$, it follows that $X^{(0)} = \{X_t^{(f)} - F(t),\ t \in [0, T]\}$. Consequently, asking a question about the first passage time of the original process $X^{(0)}$ (defined by eq. (7.27)) to some time-dependent boundary $L_t^* := L^* - F(t)$ can be re-formulated as:

$$\inf\{t \geq 0 : X^{(0)} \geq L_t^*\} = \inf\{t \geq 0 : X_t^{(f)} - F(t) \geq L^* - F(t)\} = \inf\{t \geq 0 : X_t^{(f)} \geq L^*\},$$

which is exactly in the form considered in this section.

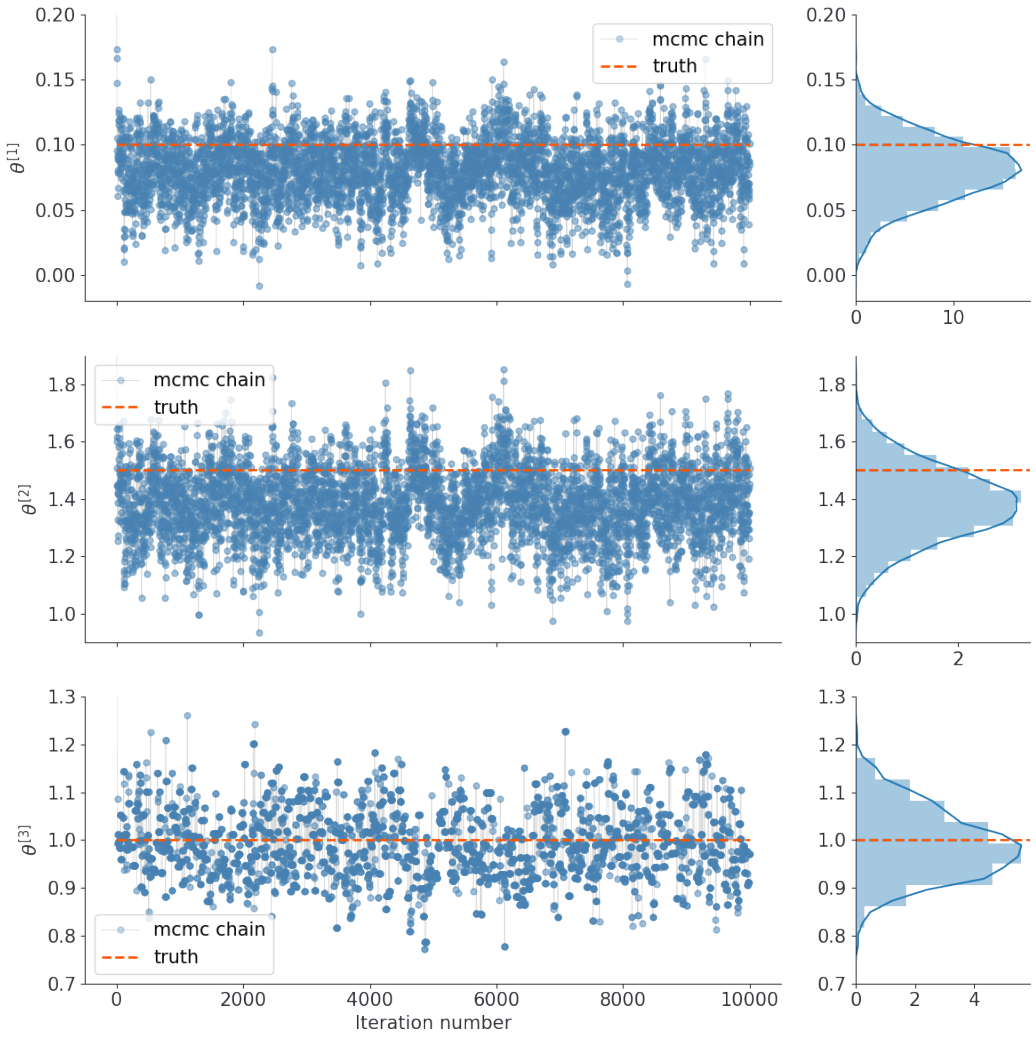


Figure 7.9: Traceplots for the inference on the modified Ornstein-Uhlenbeck process from first passage time observations. The θ chain was initialised at $\theta^{(0)} = (0.4, 3, 3)$.

7.4.1.3 Langevin diffusion

Let eq. (7.2) take the following form

$$\begin{aligned} \mathrm{d}X_t &= -\frac{(\theta^{[1]}+1)(x-\theta^{[2]})}{2[\theta^{[1]}+(x-\theta^{[2]})^2]}\,\mathrm{d}t + \theta^{[3]}\,\mathrm{d}W_t, \quad X_0 = L_*, \quad t \in [0,\tau], \\ &\text{where} \quad \tau := \inf\{t \geq 0 : X_t \geq L^*\}. \end{aligned} \tag{7.28}$$

When $\theta^{[3]} = 1$, then the process X above describes a Langevin diffusion that has t-distribution centred at $\theta^{[2]}$ with $\theta^{[1]}$ degrees of freedom as its invariant measure. The diffusion has similar properties to the Ornstein-Uhlenbeck process (indeed, as $\theta^{[1]} \to \infty$ the t-distribution approaches the Gaussian distribution); however, the

tails of the diffusion from eq. (7.28) are lighter. Because of the problems related to intractability issues, many diffusion models might have been left out from consideration in the literature on inference from first passage time observations, despite the fact that they might provide better fits. The process defined by eq. (7.28) is just one example that could be explored with the methodology introduced in this chapter that could not have been considered before. The results of the two inference procedures—one, jointly estimating $(\theta^{[1]}, \theta^{[3]})$, another, doing so for $(\theta^{[2]}, \theta^{[3]})$ are given in appendix B.

7.4.2 FitzHugh-Nagumo model

I assume that the underlying process, which I temporarily call V, solves the following stochastic differential equation:

$$\begin{aligned} \mathrm{d}V^{[1]} &= \frac{1}{\vartheta^{[1]}}\left(V_t^{[1]} - \left(V_t^{[1]}\right)^3 - V_t^{[2]} + \vartheta^{[2]}\right)\mathrm{d}t, \\ \mathrm{d}V^{[2]} &= \left(\vartheta^{[3]}V^{[1]} - V_t^{[2]} + \vartheta^{[4]}\right)\mathrm{d}t + \vartheta^{[5]}\mathrm{d}W_t, \quad V_0 = v_0, \quad t \in [0, T]. \end{aligned} \tag{7.29}$$

It can be shown that eq. (7.29) describes a hypoelliptic diffusion (Ditlevsen and Samson, 2019). Equation (7.29) gives the form—most commonly appearing in the literature–of the stochastic differential equation describing the FitzHugh-Nagumo model. However, this is not a form for which guided proposals can be rigorously applied to and thus it is imperative to transform it into the form asserted in section 7.1.3. This can be achieved by defining a simple transformation

$$f_\vartheta(v) := \begin{pmatrix} v \\ \left[v^{[1]} - \left(v^{[1]}\right)^3 - v^{[2]} + \vartheta^{[2]}\right]/\vartheta^{[1]} \end{pmatrix},$$

and associating the FitzHugh-Nagumo model with the process $X := \{f_\vartheta(V_t); t \in [0, T]\}$ instead. Process X solves the following stochastic differential equation:

$$\begin{aligned} \mathrm{d}X_t^{[1]} &= X_t^{[2]}\mathrm{d}t, \\ \mathrm{d}X_t^{[2]} &= \frac{1}{\vartheta^{[1]}}\left[\left(1-\vartheta^{[3]}\right)X_t^{[1]} + \left(1 - 3\left(X_t^{[1]}\right)^2 - \vartheta^{[1]}\right)X_t^{[2]} - \left(X_t^{[1]}\right)^3 + \vartheta^{[2]} - \vartheta^{[4]}\right]\mathrm{d}t \\ &\qquad + \frac{\vartheta^{[5]}}{\vartheta^{[1]}}\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T]. \end{aligned}$$

Additionally, for the reasons I discuss in section 7.4.2.2 below, related to the possibility of employing conjugate updates, I additionally re-parametrise the model

via:

$$\vartheta \to \theta := \left(\frac{1}{\vartheta^{[1]}}, \frac{\vartheta^{[2]}}{\vartheta^{[1]}}, \frac{\vartheta^{[3]}}{\vartheta^{[1]}}, \frac{\vartheta^{[4]}}{\vartheta^{[1]}}, \frac{\vartheta^{[5]}}{\vartheta^{[1]}} \right),$$

so that the underlying process X is defined as a solution to the SDE:

$$\begin{aligned} \mathrm{d}X_t^{[1]} &= X_t^{[2]}\,\mathrm{d}t, \\ \mathrm{d}X_t^{[2]} &= \left[\left(\theta^{[1:4]}\right)^T, 1 \right] \psi(X_t)\,\mathrm{d}t + \theta^{[5]}\,\mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T], \\ &\text{where } \psi(x) := \left(x^{[1]} - \left(x^{[1]}\right)^3 + x^{[2]} - 3\left(x^{[1]}\right)^2 x^{[2]}, 1, -x^{[1]}, -1, -x^{[2]} \right)^T. \end{aligned} \tag{7.30}$$

I simulated $K = 20$ observations from the SDE given by eq. (7.30), with the parameters set to $\theta = (10, -5, 10, 0, 3)^T$ and $(L_*, L^*) = (-0.5, 0.5)$. The diffusion was initialised at $x_0 = (-0.5, -0.975)$.

7.4.2.1 Auxiliary law

For the imputation step, I used guided proposals with the auxiliary diffusion chosen separately for each sub-interval $[\tau^*_{i-1}, \tau^*_i]$, $(i = 1, \ldots, K)$ by linearising the target SDE in eq. (7.30) at the terminal observation of this interval. More precisely, the auxiliary diffusion $\widetilde{X}$ on an interval $[\tau^*_{i-1}, \tau^*_i]$, $(i = 1, \ldots, K)$ was chosen to be the solution to:

$$\begin{aligned} \mathrm{d}\widetilde{X}_t &= \left(B_\theta \widetilde{X}_t + \beta_\theta \right) \mathrm{d}t + \begin{pmatrix} 0 \\ \theta^{[5]} \end{pmatrix} \mathrm{d}W_t, \\ &\text{where} \quad B_\theta := \begin{pmatrix} 0 & 1 \\ \theta^{[1]}\left[1 - 3(L^*)^2\right] - \theta^{[3]} & \theta^{[1]}\left[1 - 3(L^*)^2\right] - 1 \end{pmatrix} \\ &\text{and} \quad \beta_\theta := \begin{pmatrix} 0 \\ 2(L^*)^3 + \theta^{[2]} - \theta^{[4]} \end{pmatrix}. \end{aligned}$$

7.4.2.2 Conjugate updates

It is always possible to complete the step of parameter update by employing the Metropolis-Hastings algorithm based on the expression for the joint density from eq. (7.26). However, in the case of FitzHugh-Nagumo model this happens to be a sub-optimal strategy. Notice from eq. (7.30), that the drift is a simple, linear function of the parameter vector $\theta^{[1:4]}$. This makes it possible to employ conjugate updates, which I summarise in theorem 7.4.1 below.

Theorem 7.4.1. Let ϑ be some parameter of interest and let X be the solution to a two-dimensional, hypoelliptic stochastic differential equation of the form :

$$\mathrm{d}X_t = b_\vartheta(X_t)\,\mathrm{d}t + \begin{pmatrix} 0 \\ \sigma(X_t) \end{pmatrix} \mathrm{d}W_t, \quad X_0 = x_0, \quad t \in [0, T],$$

where $b_\vartheta^{[1]} \equiv b^{[1]}$ is independent of ϑ and $b_\vartheta^{[2]}$ is a linear function of ϑ, i.e. there exist a function $\psi_1(x) : \mathbb{R}^2 \to \mathbb{R}^{\bar{d}}$, where $\bar{d} := dim(\vartheta)$ and a function $\psi_2(x) : \mathbb{R}^2 \to \mathbb{R}$, both independent of ϑ, such that:

$$b_\vartheta^{[2]}(x) = \vartheta^T \psi_1(x) + \psi_2(x).$$

Define:

$$\mu^{[i]} := \int_0^T \frac{\psi_1^{[i]}(X_t)}{\sigma^2(X_t)}\,\mathrm{d}X_t^{[2]} - \int_0^T \frac{\psi_1^{[i]}(X_t)}{\sigma^2(X_t)}\psi_2(X_t)\,\mathrm{d}t, \quad i = 1, \ldots, \bar{d}$$

$$\mathscr{W}^{[i,j]} := \int_0^T \frac{\psi_1^{[i]}(X_t)\psi_1^{[j]}(X_t)}{\sigma^2(X_t)}\,\mathrm{d}t, \quad i = 1, \ldots, \bar{d}, \quad j = 1, \ldots, \bar{d}.$$

Then, the likelihood function for ϑ is conjugate to a Gaussian prior $\mathtt{Gsn}(\mu_{\mathrm{pr}}, \Sigma_{\mathrm{pr}})$, and the posterior over ϑ (conditioned on the entire path X) takes the form:

$$\vartheta | X \sim \mathtt{Gsn}\Big[\big(\mathscr{W} + \Sigma_{\mathrm{pr}}^{-1}\big)^{-1}\big(\mu + \Sigma_{\mathrm{pr}}^{-}1\mu_{\mathrm{pr}}\big), \big(\mathscr{W} + \Sigma_{\mathrm{pr}}^{-1}\big)^{-1}\Big].$$

Consequently, updates of any subset of the parameters $\theta^{[1:4]}$ can be performed with efficient conjugate updates and only the updates of $\theta^{[5]}$ need to be done with the Metropolis-Hastings algorithm based on eq. (7.26).

7.4.2.3 Inference results

I performed joint inference of the parameters $\theta^{[2]}$ and $\theta^{[3]}$. I used independent Gaussian priors with mean 0 and variance 1000. The starting point was assumed to be unobserved. I put an independent Gaussian prior with mean 0 and variance 100 on each coordinate of the starting point. I used 7th order 7/6 Runge-Kutta method (Verner, 1978) for solving ODEs that determine $\widetilde{H}$, $\widetilde{r}$ and $\widetilde{h}$ terms. I also started with an equidistant time-grid with a step size 0.002 that I subsequently transformed according to the usual guidelines (see remark 6.2.2). I used the preconditioned Crank-Nicolson scheme with memory parameter $\lambda = 0.997$, but did not use any blocking. The MCMC chain was initiated at $\theta^{[2;3](0)} = (-5, 10)^T$ and it

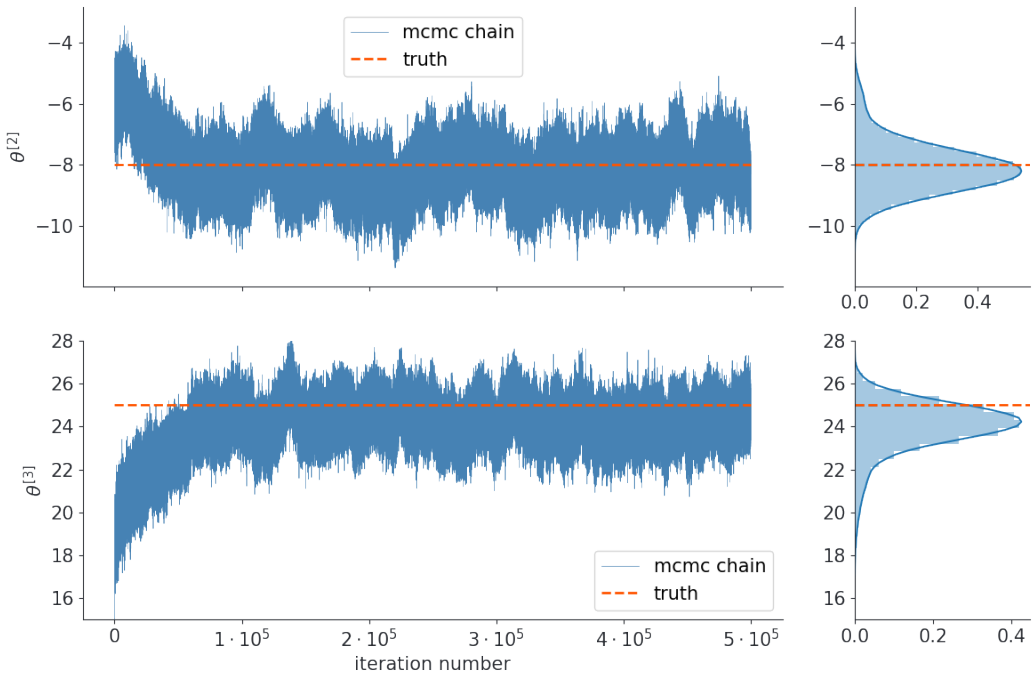


Figure 7.10: Results of inference from first passage time observations for the FitzHugh-Nagumo model.

was run for $5 \cdot 10^5$ iterations. Figure 7.10 summarises the results of this inference run. Even at such a small sample size, the joint posterior over both parameters have been successfully identified. Additionally, in fig. 7.11 I plot a sample of paths that were imputed during the inference. The trajectories were transformed back to the original parametrisation in eq. (7.29). Notice that only the points marked with dots are the observations, everything else, including the entire second coordinate of the trajectory, as well as the starting point, was unobserved.

7.5 Discussion

In this chapter I presented a detailed treatment of the problem of Bayesian inference for diffusion processes from first passage time observations. I considered one dimensional, stochastic leaky integrate-and-fire models, multidimensional uniformly elliptic models and multidimensional hypoelliptic models.

For the former ones I introduced a number of imputation-based inference algorithms that provide solutions for a much broader class of diffusion processes and work under a wider array of experimental designs than the current state-of-the-art

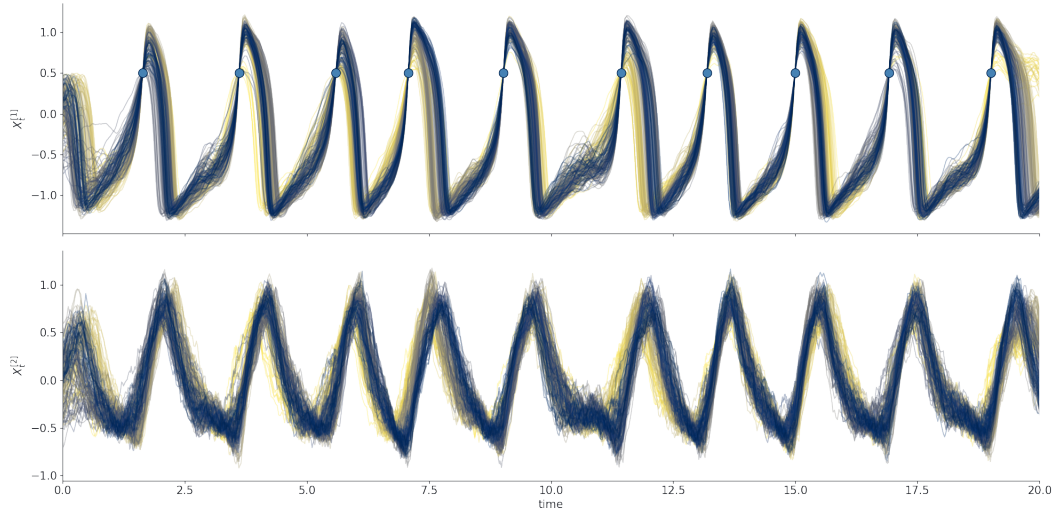


Figure 7.11: Paths of the FitzHugh-Nagumo model that were imputed during inference. The trajectories were transformed back to the original parametrisation in eq. (7.29).

methodologies do. Additionally, one of the introduced procedures allows for inference devoid of any discretisation errors. In appendix A I also showed how to naturally extend some of those algorithm to target multidimensional, uniformly elliptic diffusions.

For the hypoelliptic models I introduced an inference algorithm that effectively allows to employ the version of guided proposals suitable for partially observed diffusions by appending an additional rejection step.

In the numerical section I illustrated the flexibility of the introduced methodologies by considering intractable diffusion processes and composite observational regimes. Additionally, I illustrated that even the complex and highly relevant FitzHugh-Nagumo model can be tackled with the introduced algorithms.

There are a number of extensions to models in neuroscience that could be handled by modified versions of the algorithm presented in this chapter. For instance, the assumption about a constant firing threshold L^* could be lifted and substituted with a height-dependent firing intensity. Research on this and other extensions relevant for neuroscience is ongoing.

Throughout, I focused on the applications to neuroscience; however, there are other fields that can immediately benefit from the results presented here. In particular, extensions to some problems from particle physics and population dynamics are currently being researched.

Proofs

Proof of theorem 7.2.1. Following the argument from section 2.3.3 (see also Roberts and Stramer (2001)), taking $\mathcal{Z} := \tau$ and $\varrho(\mathrm{d}\mathbb{z})$ to be the Lebesgue measure I have the following factorisation:

$$\mathbb{P}_0^{(\theta)} \cdot \mathbb{1}_{\tau<\infty} = \mathbb{P}_0^{(\theta)}(\cdot|\tau) \otimes g_\mu^{(\theta)}(\tau)\varrho(\mathrm{d}\tau)\mathbb{1}_{\tau<\infty}, \quad \mathbb{P}_\alpha^{(\theta)} \cdot \mathbb{1}_{\tau<\infty} = \mathbb{P}_\alpha^{(\theta)}(\cdot|\tau) \otimes g_\alpha^{(\theta)}(\tau)\varrho(\mathrm{d}\tau)\mathbb{1}_{\tau<\infty},$$

from which it follows that

$$\frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(Y|\mathcal{Z}) = \frac{g_0^{(\theta)}(\tau)}{g_\alpha^{(\theta)}(\tau)} \frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(Y).$$

The statement of the theorem follows after deriving $[\mathrm{d}\mathbb{P}_\alpha^{(\theta)} \big/ \mathrm{d}\mathbb{P}_0^{(\theta)}](Y)$, which can be done by adapting calculations from section 2.3 in a straightforward way so as to accommodate time-dependent coefficients. □

Proof of lemma 7.2.1. Write $\mathbb{Q}(\cdot) := \mathbb{P}_b^{(\theta)}(\cdot|X_{\tau_1}^{[1]} = L^*)$, for some fixed $\tau_1 \in \mathbb{R}_+$. Then, for $\epsilon \in (0, \tau_1)$:

$$\begin{aligned}
\mathbb{P}_b^{(\theta)}(\{\omega \in \Omega : \tau_1^*(\omega) = \tau_1\}|X_{\tau_1}^{[1]} = L^*) &= \mathbb{Q}\left(X_t^{[1]} < L^*, \forall t \in [0, \tau_1)\right) \\
&= \int \mathbb{Q}\left(X_t^{[1]} < L^*, \forall t \in [0, \tau_1)\middle| X_{\tau_1-\epsilon} = x\right) \\
&\qquad \cdot \mathbb{Q}(X_{\tau_1-\epsilon} \in \mathrm{d}x) \\
&= \int \mathbb{Q}\left(X_t^{[1]} < L^*, \forall t \in [0, \tau_1-\epsilon)\middle| X_{\tau_1-\epsilon} = x\right) \\
&\quad \cdot \mathbb{Q}\left(X_t^{[1]} < L^*, \forall t \in [\tau_1-\epsilon, \tau_1)\middle| X_{\tau_1-\epsilon} = x\right) \\
&\quad \cdot f_{L^*}(0, x_0, \tau_1-\epsilon, x)\,\mathrm{d}x,
\end{aligned}$$

where $f_{L^*}(0, x_0, t, x)$ denotes the transition density of X under the law $\mathbb{Q}$:

$$f_{L^*}(0, x_0, t, x)\,\mathrm{d}x := \mathbb{Q}(X_t \in \mathrm{d}x), \quad t \in [0, \tau_1), \quad x \in (-\infty, L^*) \times \mathbb{R}^{d-1},$$

(which is well-defined by the virtue of X being hypoelliptic), and where conditioning on $X_{\tau_1-\epsilon} = x$ granted the use of Markov property. Define $\mathscr{R}_\epsilon := \{x \in \mathscr{X} : b^{[1]}(x) > \epsilon\}$ and $\mathscr{H}_\delta := \{x \in (-\infty, L^* - \delta] \times \mathbb{R}^{d-1}\}$, and notice that in order for τ^* to be well defined, there must exist an $\epsilon > 0$ and $\delta > 0$ for which $\mathscr{R}_\epsilon \cap \mathscr{H}_\delta$ has a positive mass under the target measure $\mathbb{P}_b^{(\theta)}$. Indeed, since the function b is continuous, if there were no $\epsilon > 0$ and $\delta > 0$ for which $\mathscr{R}_\epsilon \cap \mathscr{H}_\delta$ had positive mass under

the target measure, then whenever $X_0 \in \bigcup_{\delta>0} \mathscr{H}_\delta$: $\mathrm{d}X_t^{[1]} := b(X_t)\mathrm{d}t \leq 0$ with probability one and consequently, with probability one, $X^{[1]}$ would have never reached L^*. Fix ϵ and δ to such values. It now follows that

$$\begin{aligned}\mathbb{P}_b^{(\theta)}(\{\omega\in\Omega\colon \tau_1^*(\omega)=\tau_1\}|X_{\tau_1}^{[1]}=L^*) \geq \int_{\mathscr{H}_\delta\cap\mathscr{R}_\epsilon} &\mathbb{Q}\left(X_t^{[1]}<L^*, \forall t\in[0,\tau_1-\epsilon)\middle|X_{\tau_1-\epsilon}=x\right)\\ &\cdot\mathbb{Q}\left(X_t^{[1]}<L^*, \forall t\in[\tau_1-\epsilon,\tau_1)\middle|X_{\tau_1-\epsilon}=x\right)\\ &\cdot f_{L^*}(0,x_0,\tau_1-\epsilon,x)\mathrm{d}x>0,\end{aligned}$$

where I used that by the definitions of $\mathscr{H}_\delta$ and $\mathscr{R}_\epsilon$ (and hypoellipticity of X, implying smoothness of the transition densities):

$$\mathbb{Q}\left(X_t^{[1]}<L^*, \forall t\in[0,\tau_1-\epsilon)\middle|X_{\tau_1-\epsilon}=x\right)>0, \quad \text{on } x\in\mathscr{H}_\delta,$$

and

$$\mathbb{Q}\left(X_t^{[1]}<L^*, \forall t\in[\tau_1-\epsilon,\tau_1)\middle|X_{\tau_1-\epsilon}=x\right)>0, \quad \text{on } x\in\mathscr{R}_\epsilon\cap\mathscr{H}_\delta.$$

□

Proof of theorem 7.2.2. The result follows immediately from lemma 7.2.1 and Bayes' theorem. □

Proof of theorem 7.3.1. This proof follows by directly adapting the proof of Sermaidis et al. (2013, Theorem 1). There are three differences I would like to point out. First, unlike Sermaidis et al. (2013), who take $\mathcal{Z}_i$ to be the conditioning on the exact observation of the diffusion X at its terminal point, which leads to appearance of the term $g_0(\mathcal{Z}_i)$ (representing the transition density under the Wiener measure) I take $\mathcal{Z}_i$ to represent first passage time to level L^* and this leads to appearance of the term $J_{L^*\theta-L_*\theta}(\tau_i)$ instead (representing the first passage time density under the Wiener measure). Second, because $g_0(\mathcal{Z}_i)$ concerns the state variable and the diffusion X is transformed via Lamperti transformation to diffusion Y an additional term appears in the density of Sermaidis et al. (2013, Theorem 1) (the determinant of a Jacobian of the Lamperti transformation). This term is no longer present in eq. (7.25), as $J_{L^*\theta-L_*\theta}(\tau_i)$ is not a function of the state variable. Finally, I introduced explicit dependence on the time variable, which is straightforward to accommodate. All other reasoning remains unchanged. □

Proof of theorem 7.3.2. This follows by the standard arguments, for instance from Beskos and Roberts (2005), appropriately modified to accommodate time-dependent terms. For completeness I write out the argument. If the pair (B,θ) are such that $\mathscr{D}$ occur for a trajectory $\Psi_\theta(B)$, then:

$$\begin{aligned}\pi(\theta,B,\mathscr{D}) &= \pi(\theta)\pi(\mathscr{D}|\theta)\pi(B|\mathscr{D},\theta)\\ &= \pi(\theta)g_\alpha^{(\theta)}(\mathscr{D})\frac{g_0^{(\theta)}}{g_\alpha^{(\theta)}}(\mathscr{D})\frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(\Psi_\theta(B))\\ &= \pi(\theta)g_0^{(\theta)}(\mathscr{D})\prod_{i=1}^{K}\exp\Big\{A(\nu_i+\tau_i,L_\theta^{*[i]})-A(\nu_i,L_{*\theta}^{\ [i]})\\ &\qquad -\int_0^{\tau_i}\varphi_\theta\Big[\nu_i+t,\Psi_\theta^{[i]}(B^{[1:3,i]})_t\Big]\,\mathrm{d}t\Big\}.\end{aligned}$$

Now, $g_0^{(\theta)}(\mathscr{D}) := \prod_{i=1}^K g_0^{(\theta)}(\tau_i)$, where each $g_0^{(\theta)}(\tau_i)$ is the density for the first passage time of Brownian motion (started from $L_{*\theta}^{\ [i]}$) to the level $L_\theta^{*[i]}$, that is evaluated at the time τ_i, and it is given in eq. (7.16). Otherwise, if path $\Psi_\theta(B)$ does not go through $\mathscr{D}$, the density vanishes. □

Proof of theorem 7.3.3. Using that $\mathcal{Z}\equiv\mathscr{D}$ and $\mathcal{Z}\subset\widehat{\mathcal{Z}}$:

$$\pi(\theta,W,\mathscr{D}) = \pi(\theta,W,\mathcal{Z}) = \pi(\theta,W,\mathcal{Z},\widehat{\mathcal{Z}}) = \pi(\theta)\pi(\widehat{\mathcal{Z}}|\theta)\pi(\mathcal{Z}|\widehat{\mathcal{Z}},\theta)\pi(W|\mathcal{Z},\theta).$$

Now, denoting by $\widehat{h}_\theta(0,x_0) := \mathrm{d}\mathbb{P}_b^{(\theta)}(\widehat{\mathcal{Z}})$, I have by eq. (7.20) and eq. (3.24):

$$\begin{aligned}\pi(\theta,W,\mathscr{D}) &= \pi(\theta)\widehat{h}_\theta(0,x_0)\mathbb{P}_b^{(\theta)}(\mathcal{Z}|\widehat{\mathcal{Z}})\frac{\mathbb{1}_{\mathcal{Z}}(\Psi_\theta(W))}{\mathbb{P}_b^{(\theta)}(\mathcal{Z}|\widehat{\mathcal{Z}})}\frac{\mathrm{d}\mathbb{P}_b^{(\theta)}(\Psi_\theta(W)|\widehat{\mathcal{Z}})}{\mathrm{d}\mathbb{P}_{b^\circ}^{(\theta)}(\Psi_\theta(W))}\\ &= \pi(\theta)\widehat{h}_\theta(0,x_0)\frac{\widetilde{h}_\theta(0,x_0)}{\widehat{h}_\theta(0,x_0)}\exp\left\{\int_0^{\tau_K}G_\theta[\Psi_\theta(W)]\,\mathrm{d}t\right\}\mathbb{1}_{\mathcal{Z}}(\Psi_\theta(W))\\ &= \pi(\theta)\widetilde{h}_\theta(0,x_0)\exp\left\{\int_0^{\tau_K}G_\theta[\Psi_\theta(W)]\,\mathrm{d}t\right\}\mathbb{1}_{\mathscr{D}}(\Psi_\theta(W)).\end{aligned}$$

□

Proof of theorem 7.4.1. Define a two-dimensional process V as a solution to the following SDE:

$$\mathrm{d}V_t = a(V_t)\,\mathrm{d}t + \begin{pmatrix}0\\ \sigma(V_t)\end{pmatrix}\mathrm{d}W_t,\quad V_0 = x_0,\quad t\in[0,T],$$

where $a^{[1]} \equiv b^{[1]}$ and $a^{[2]} \equiv 0$. Denote by $\mathbb{Q}$ the law induced by it. Then, recall from example 2.2.1, that for a diffusion that has a form from theorem 7.4.1, the Radon-Nikodým derivative between its law and the law $\mathbb{Q}$ is given by:

$$\frac{\mathrm{d}\mathbb{P}_b^{(\vartheta)}}{\mathrm{d}\mathbb{Q}}(X) = \exp\left\{\int_0^T \left[b_\vartheta^{[2]}/\sigma^2\right](X_s)\,\mathrm{d}X_s^{[2]} - \frac{1}{2}\int_0^T \left[b_\vartheta^{[2]}/\sigma\right]^2(X_s)\,\mathrm{d}s\right\}.$$

Using the $(\mu, \mathscr{W})$ notation, the above can be re-written as:

$$\frac{\mathrm{d}\mathbb{P}_b^{(\vartheta)}}{\mathrm{d}\mathbb{Q}}(X) = \exp\left\{\vartheta^T\mu - \frac{1}{2}\vartheta^T\mathscr{W}\vartheta + R\right\},$$

where

$$R := \int_0^T \psi_2(X_t)/\sigma^2(X_t)\,\mathrm{d}X_t^{[2]} - \frac{1}{2}\int_0^T \left[\vartheta^T\psi_2(X_t)\right]^2/\sigma^2(X_t)\,\mathrm{d}t.$$

The conditioned Radon-Nikodým derivative becomes:

$$\frac{\mathrm{d}\mathbb{P}_b^{(\vartheta)}}{\mathrm{d}\mathbb{Q}}(X|\mathscr{D}) = \frac{q(\mathscr{D})}{g_b^{(\vartheta)}(\mathscr{D})}\exp\left\{\vartheta^T\mu - \frac{1}{2}\vartheta^T\mathscr{W}\vartheta + R\right\},$$

where $q(\mathscr{D})$ is the density for observing the data under the measure $\mathbb{Q}$, and where $g_b^{(\vartheta)}(\mathscr{D})$ is the corresponding density under the target measure $\mathbb{P}_b^{(\vartheta)}$. It now follows that the joint density over the parameters, path and the data is given by:

$$\begin{aligned}\pi(\vartheta, X, \mathscr{D}) &= \pi(\vartheta)\pi(\mathscr{D}|\vartheta)\pi(X|\vartheta,\mathscr{D}) \\ &= \pi(\vartheta)q(\mathscr{D})\exp\left\{\vartheta^T\mu - \frac{1}{2}\vartheta^T\mathscr{W}\vartheta + R\right\}.\end{aligned}$$

Finally, deriving $\pi(\vartheta|\mathscr{D}, X)$ for a choice of prior from the theorem follows readily after completing the square for Gaussian densities. □

A Appendix

A.1 Multidimensional, uniformly elliptic models

Just as in section 7.2.3, I denote by $\mathbb{P}_b^{(\theta)}$ the unconditioned law induced by eq. (7.3); notice that the conditioned-on variable, corresponding to the dataset (7.5) is given by:

$$\mathscr{Z} := \{\tau_i^*\,;\, i = 1, \ldots, K\}.$$

For simplicity of exposition I assume that $K = 2$, so that:

$$\mathcal{Z} := \{\tau_1^*, \tau_2^*\},$$

where the general case will follow immediately by induction.

A.1.1 Sampling conditioned diffusion paths

Throughout, I assume that first passage times of the first coordinate $X^{[1]}$ are observed. The target law is not factorised by the stopping times τ_i^*, $(i = 1,\dots,K)$ into independent laws defined on shorter time segments, and thus global proposals need to be found.

To this end, define:

$$\widehat{\mathcal{Z}} := \mathcal{Z} \cup \{\tau_{*1}\} = \{\tau_1^*, \tau_{*1}, \tau_2^*\},$$

and consider a simpler problem of sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$. It is the law of the target diffusion (7.3), conditioned on the first up-crossing time τ_1^*, subsequent down-crossing time τ_{*1} and a final up-crossing time τ_2^*. Assume A1, A3, A5–A8 and A10 hold. Additionally assume:

Assumption A18. *The first coordinate of the image of the Lamperti transformation $\eta_\theta^{[1]}(\cdot)$ does not depend on the latent coordinates $X^{[2:d]}$. In particular, for any $x \in \mathbb{R}$ there exist $c_x \in \mathbb{R}$ such that: $\eta_\theta^{[1]}((x, e_1^T)^T) = \eta_\theta^{[1]}((x, e_2^T)^T) = c_x$, for all $e_1, e_2 \in \mathbb{R}^{d-1}$.*

First, using Lamperti transformation η_θ define the process Y as: $Y := \{\eta_\theta(X_t),\, t \in [0, T]\}$, which solves

$$\mathrm{d}Y_t = \alpha_\theta(Y_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad Y_0 = y_0 =: \eta_\theta(x_0), \quad t \in [0, T].$$

The stopping times $(\tau_1^*, \tau_{*1}, \tau_2^*)$ may be expressed in terms of the process Y as follows:

$$\begin{cases} \tau_1^*(\theta) = \inf\{t \geq 0 : Y_t^{[1]} = L^*{}_\theta\}, \\ \tau_{*1}(\theta) = \inf\{t \geq \tau_1^*(\theta) : Y_t^{[1]} = L_{*\theta}\}, \\ \tau_2^*(\theta) = \inf\{t \geq \tau_{*1}(\theta) : Y_t^{[1]} = L^*{}_\theta\}, \end{cases}$$

where $L_{*\theta} := \eta_\theta^{[1]}((L_*, \cdot)^T)$ and $L^*{}_\theta := \eta_\theta^{[1]}((L^*, \cdot)^T)$.

Consider the proposal law $\mathbb{P}_0^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$, given by the d-dimensional Wiener measure conditioned on $\widehat{\mathcal{Z}}$. By the same token as theorem 7.2.1, the Radon-Nikodým derivative between the target and the proposal laws is given by:

$$\frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(Y|\widehat{\mathcal{Z}}) = \frac{g_0^{(\theta)}(\widehat{\mathcal{Z}})}{g_\alpha^{(\theta)}(\widehat{\mathcal{Z}})} \exp\left\{\int_0^{\tau_2} \alpha_\theta^T(Y_s)\mathrm{d}Y_s - \frac{1}{2}\int_0^{\tau_2} ||\alpha_\theta(Y_s)||_2^2 \mathrm{d}s\right\}, \tag{7.31}$$

where I used the statement of the Girsanov theorem from eq. (2.3), instead of eq. (2.8) that was used in the proof of theorem 7.2.1. $g_0^{(\theta)}(\widehat{\mathcal{Z}})$ (resp. $g_\alpha^{(\theta)}(\widehat{\mathcal{Z}})$) denotes the density under the proposal (resp. target) measure for observing $\widehat{\mathcal{Z}}$.

To make use of eq. (7.31), I can sample from the proposal law $\mathbb{P}_0^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$ and correct for the deviation from the target law using the Radon-Nikodým derivative above. Under the proposal law, coordinates of Y are independent and thus $(\tau_1^*, \tau_{*1}, \tau_2^*)$ factorise the law of the first coordinate $Y^{[1]}$ into three independent laws defined on the time-segments $[0, \tau_1^*]$, $[\tau_1^*, \tau_{*1}]$ and $[\tau_{*1}, \tau_1^*]$. Define:

$$Z_t^{[1]} := L^*{}_\theta - Y^{[1]}_{\tau_1^* - t}, \quad t \in [0, \tau_1^*], \qquad Z_t^{[2]} := Y^{[1]}_{\tau_{*1} - t} - L_{*\theta}, \quad t \in [0, \tau_{*1} - \tau_1^*],$$
$$Z_t^{[3]} := L^*{}_\theta - Y^{[1]}_{\tau_2^* - t}, \quad t \in [0, \tau_2^* - \tau_{*1}].$$

By the same arguments as the ones used in section 7.2.1.1, it follows that if Y is a sample from the proposal law $\mathbb{P}_0^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$ then $Z^{[k]}$, $k = 1, 2, 3$ are three independent Bessel bridges that admit decomposition in eq. (7.13) and thus can be simulated using, altogether, nine independent Brownian bridges. Notice that the remaining coordinates $\{Y_t^{[2:d]},\ t \in [0, \tau_2^*]\}$ are distributed jointly as a $(d-1)$-dimensional Brownian motion started from $y_0^{[2:d]}$.

Neither $g_\alpha^{(\theta)}(\widehat{\mathcal{Z}})$ nor $g_0^{(\theta)}(\widehat{\mathcal{Z}})$ has to be known to complete the step of path imputation; however, the latter density will be essential for the step of updating unknown variables. Similarly to the explanation in remark 7.2.1, it follows directly from the definition of the proposal measure and the strong Markov property that $g_0^{(\theta)}(\widehat{\mathcal{Z}})$ can be expressed as a product of three first passage time densities of Brownian motion:

$$g_0^{(\theta)}(\widehat{\mathcal{Z}}) = J_{L^*{}_\theta - y_0^{[1]}}(\tau_1^*) J_{L^*{}_\theta - L_{*\theta}}(\tau_* - \tau_1^*) J_{L^*{}_\theta - L_{*\theta}}(\tau_2^* - \tau_{*1}). \tag{7.32}$$

For the reasons that will shortly become apparent, take the non-centred process to be given by nine independent 0–0 Brownian bridges on $[0, 1]$: $B := (B^{[i]})_{i=1}^9$, together with a $(d-1)$-dimensional, standard Brownian motion on $[0, 1]$: W (notice

that (B, W) is not only independent from θ, but also from τ_1^*, τ_{*1} and τ_2^*). Then, define function $\Psi_{\theta,\tau_{*1}}$ by specifying separately what is done to the first coordinate and the remaining $(d-1)$ coordinates:

$$\begin{aligned}\Psi^{[1]}_{\theta,\tau_{*1}}(B,W) := &\left\{L^*{}_\theta - \widehat{\Psi}_\theta\left(B^{[1:3]}, t, \tau_1^*, \tau_1^*\right),\ t \in [0, \tau_1^*]\right\} \\ &\cup \left\{\widehat{\Psi}_\theta\left(B^{[4:6]}, t, \tau_{*1}, \tau_{*1} - \tau_1^*\right) - L_{*\theta},\ t \in [\tau_1^*, \tau_{*1}]\right\} \\ &\cup \left\{L^*{}_\theta - \widehat{\Psi}_\theta\left(B^{[7:9]}, t, \tau_2^*, \tau_2^* - \tau_{*1}\right),\ t \in [\tau_{*1}, \tau_2^*]\right\}, \quad \text{where}\end{aligned}$$

$$\widehat{\Psi}_\theta(x, t, \tau, \bar{\tau}) := \sqrt{\left(\sqrt{\bar{\tau}} x^{[1]}_{(\tau-t)/\bar{\tau}} + (L^*{}_\theta - L_{*\theta})\frac{\tau - t}{\bar{\tau}}\right)^2 + \bar{\tau}\left(x^{[2]}_{(\tau-t)/\bar{\tau}}\right)^2 + \bar{\tau}\left(x^{[3]}_{(\tau-t)/\bar{\tau}}\right)^2},$$

$$\Psi^{[2:d]}_{\theta,\tau_{*1}}(B,W) := \left\{y_0^{[2:d]} + \sqrt{\tau_2^*} W_{t/\tau_2^*};\ t \in [0, \tau_2^*]\right\}.$$

In particular, $\Psi_{\theta,\tau_{*1}}(B,W) \sim \mathbb{P}_0^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$.

The above gives a recipe for sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$. Recall however that, in practice, τ_{*1} is latent and it is an algorithm for sampling from $\mathbb{P}_b^{(\theta)}(\cdot|\mathcal{Z})$ that is sought after. To address this, I define a Gibbs sampler which alternately samples from $\pi(B, W|\tau_{*1}, \mathcal{Z})$ and $\pi(\tau_{*1}|B, W, \mathcal{Z})$. The resulting chain of $\{\Psi_{\theta,\tau_{*1}^{(n)}}(B^{(n)}, W^{(n)});\ n = 1, \ldots, N\}$ has $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$ as its invariant measure. The former update can be completed with a non-centrally parametrised version of the Metropolis-Hastings algorithm targeting law $\mathbb{P}_\alpha^{(\theta)}(\cdot|\widehat{\mathcal{Z}})$ that has been described above. For the latter step, notice that:

$$\begin{aligned}\pi(\tau_{*1}|B, W, \mathcal{Z}) &\propto \pi(\tau_{*1}, \mathcal{Z})\pi(B, W|\tau_{*1}, \mathcal{Z}) \\ &= g_\alpha^{(\theta)}(\widehat{\mathcal{Z}})\frac{\mathrm{d}\mathbb{P}_\alpha^{(\theta)}}{\mathrm{d}\mathbb{P}_0^{(\theta)}}(\Psi_{\theta,\tau_{*1}}(B,W)|\widehat{\mathcal{Z}}) \\ &= g_0^{(\theta)}(\widehat{\mathcal{Z}})\exp\left\{\int_0^{\tau_2} \alpha_\theta^T\left(\left[\Psi_{\theta,\tau_{*1}}(B,W)\right]_s\right)\mathrm{d}\left[\Psi_{\theta,\tau_{*1}}(B,W)\right]_s \right. \\ &\qquad \left. -\frac{1}{2}\int_0^{\tau_2}\left\|\alpha_\theta\left(\left[\Psi_{\theta,\tau_{*1}}(B,W)\right]_s\right)\right\|_2^2 \mathrm{d}s\right\},\end{aligned} \tag{7.33}$$

where the final equality followed from eq. (7.31). Since eq. (7.33) can be approximated with left-Riemann sums, it is possible to employ the Metropolis-Hastings algorithm for updating τ_{*1}. The entire procedure for sampling from $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$ is summarised in algorithms 7.6 and 7.7 below.

Algorithm 7.6 Metropolis-within-Gibbs sweep for sampling from $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$

Draw $W^\circ \sim \bigotimes_{i=1}^{d-1} \mathbb{W}$ ▷ $(d-1)$-dimensional Brownian motion on $[0,1]$
Draw $B^\circ \sim \bigotimes_{i=1}^{9} \mathbb{W}^*$ ▷ Nine 0–0 Brownian bridges on $[0,1]$
Set $Y^\circ \leftarrow \Psi_{\theta,\tau_{*1}^{(n-1)}}(B^\circ, W^\circ)$
Draw $E \sim \texttt{Exp(1)}$
if $E \geq -\int_0^{\tau_2} \alpha_\theta^T(Y_s^\circ)\mathrm{d}Y_s^\circ + \frac{1}{2}\int_0^{\tau_2}||\alpha_\theta(Y_s^\circ)||_2^2\mathrm{d}s + \int_0^{\tau_2}\alpha_\theta^T(Y_s^{(n-1)})\mathrm{d}Y_s^{(n-1)} - \frac{1}{2}\int_0^{\tau_2}||\alpha_\theta(Y_s^{(n-1)})||_2^2\mathrm{d}s$ **then**
    Set $(B^{(n)}, W^{(n)}) \leftarrow (B^\circ, W^\circ)$
else
    Set $(B^{(n)}, W^{(n)}) \leftarrow (B^{(n-1)}, W^{(n-1)})$
Draw $\tau_{*1}^\circ \sim q(\tau_{*1}^{(n-1)}, \cdot)$ ▷ From some proposal kernel q
Draw $U \sim \texttt{Unif([0,1])}$
if $U \leq \frac{\pi(\tau_{*1}|B^{(n)},W^{(n)},\mathcal{Z})q(\tau_{*1}^\circ,\tau_{*1}^{(n-1)})}{\pi(\tau_{*1}|B^{(n)},W^{(n)},\mathcal{Z})q(\tau_{*1}^{(n-1)},\tau_{*1}^\circ)}$ **then** ▷ See eq. (7.33)
    Set $(\tau_{*1}^{(n)}, Y^{(n)}) \leftarrow (\tau_{*1}^\circ, \Psi_{\theta,\tau_{*1}^\circ}(B^{(n)}, W^{(n)}))$
else
    Set $(\tau_{*1}^{(n)}, Y^{(n)}) \leftarrow (\tau_{*1}^{(n-1)}, \Psi_{\theta,\tau_{*1}^{(n-1)}}(B^{(n)}, W^{(n)}))$
return $(\tau_{*1}^{(n)}, B^{(n)}, W^{(n)}, Y^{(n)})$

Algorithm 7.7 Metropolis-within-Gibbs algorithm for sampling from $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$

Initialise $\tau_{*1}^{(0)}$
Draw $W^{(0)} \sim \bigotimes_{i=1}^{d-1} \mathbb{W}$ ▷ $(d-1)$-dimensional Brownian motion on $[0,1]$
Draw $B^{(0)} \sim \bigotimes_{i=1}^{9} \mathbb{W}^*$ ▷ Nine 0–0 Brownian bridges on $[0,1]$
Set $Y^{(0)} \leftarrow \Psi_{\theta,\tau_{*1}^{(0)}}(B^{(0)}, W^{(0)})$
for $n = 1, \ldots, N$ **do**
    Draw $(B^{(n)}, W^{(n)}, Y^{(n)}, \tau_{*1}^{(n)})$ as per algorithm 7.6
return $\{Y^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant law $\mathbb{P}_\alpha^{(\theta)}(\cdot|\mathcal{Z})$

Usual remarks about blocking and the preconditioned Crank-Nicolson scheme apply.

A.1.2 Inference

To perform inference it is enough to introduce a small modification to algorithm 7.7, analogously to how it was done in sections 7.3.1 and 7.3.2. To this end, define a Gibbs sampler which alternately updates parameter θ, τ_{*1} and (B, W). Updating the latter two is done via algorithm 7.6. To update the former, notice that the joint

density is given by:

$$\begin{aligned}\pi(\theta, B, W, \tau_{*1}, \mathscr{D}) &= \pi(\theta)\pi(\tau_{*1}, \mathcal{Z}|\theta)\pi(B, W|\tau_{*1}, \mathcal{Z}, \theta) \\ &= \pi(\theta) g_0^{(\theta)}(\widehat{\mathcal{Z}}) \exp\left\{ \int_0^{\tau_2} \alpha_\theta^T \left(\left[\Psi_{\theta,\tau_{*1}}(B, W) \right]_s \right) \mathrm{d}\left[\Psi_{\theta,\tau_{*1}}(B, W) \right]_s \right. \\ &\qquad \left. -\frac{1}{2} \int_0^{\tau_2} \left\| \alpha_\theta \left(\left[\Psi_{\theta,\tau_{*1}}(B, W) \right]_s \right) \right\|_2^2 \mathrm{d}s \right\},\end{aligned} \tag{7.34}$$

In view of the expression in eq. (7.32), the density in eq. (7.34) can be approxi-

Algorithm 7.8 Inference from FPT for multidimensional, uniformly elliptic diffusions

1: Initialise ${\tau_{*1}}^{(0)}$ and $\theta^{(0)}$
2: Initialise $W^{(0)}$, $B^{(0)}$ and $Y^{(0)}$ as in algorithm 7.7
3: **for** $n = 1, \ldots, N$ **do**
4: Draw $(B^{(n)}, W^{(n)}, \_, {\tau_{*1}}^{(n)})$ as per algorithm 7.6
5: Draw $\theta^\circ \sim q(\theta^{(n-1)}, \cdot)$ ▷ For some transition kernel q
6: $U \sim$ `Unif([0,1])`
7: **if** $U \leq \frac{\pi(\theta^\circ, B^{(n)}, W^{(n)}, {\tau_{*1}}^{(n)}, \mathscr{D}) q(\theta^\circ, \theta^{(n-1)})}{\pi(\theta^{(n-1)}, B^{(n)}, W^{(n)}, {\tau_{*1}}^{(n)}, \mathscr{D}) q(\theta^{(n-1)}, \theta^\circ)}$ **then** ▷ See eq. (7.34)
8: Set $(\theta^{(n)}, Y^{(n)}) \leftarrow (\theta^\circ, \Psi_{\theta^\circ, {\tau_{*1}}^{(n)}}(B^{(n)}, W^{(n)}))$
9: **else**
10: Set $(\theta^{(n)}, Y^{(n)}) \leftarrow (\theta^{(n-1)}, \Psi_{\theta^{(n-1)}, {\tau_{*1}}^{(n)}}(B^{(n)}, W^{(n)}))$
11: **return** $\{\theta^{(n)}; n = 0, \ldots, N\}$ ▷ Markov chain with invariant density $\pi(\theta|\mathcal{Z})$

mated with left-Riemann sums. Consequently, it is possible to employ Metropolis-Hastings algorithm for updating θ. The full procedure delineating inference is summarised in algorithm 7.8.

B Appendix

B.1 Inference results for a Langevin diffusion

The parameters of the stochastic differential equation in eq. (7.28) were set to $\theta = (5, 0, 1)$ and $(L_*, L^*) = (-10, -6)$ and $K = 40$ observations were simulated. Then, I conducted two separate inference procedure: one, estimating $(\theta^{[1]}, \theta^{[3]})$, another, estimating $\theta^{[2:3]}$. The results are plotted in figs. 7.12 and 7.13 respectively. The priors and transition kernels have been set in the same way as in section 7.4.1.1.

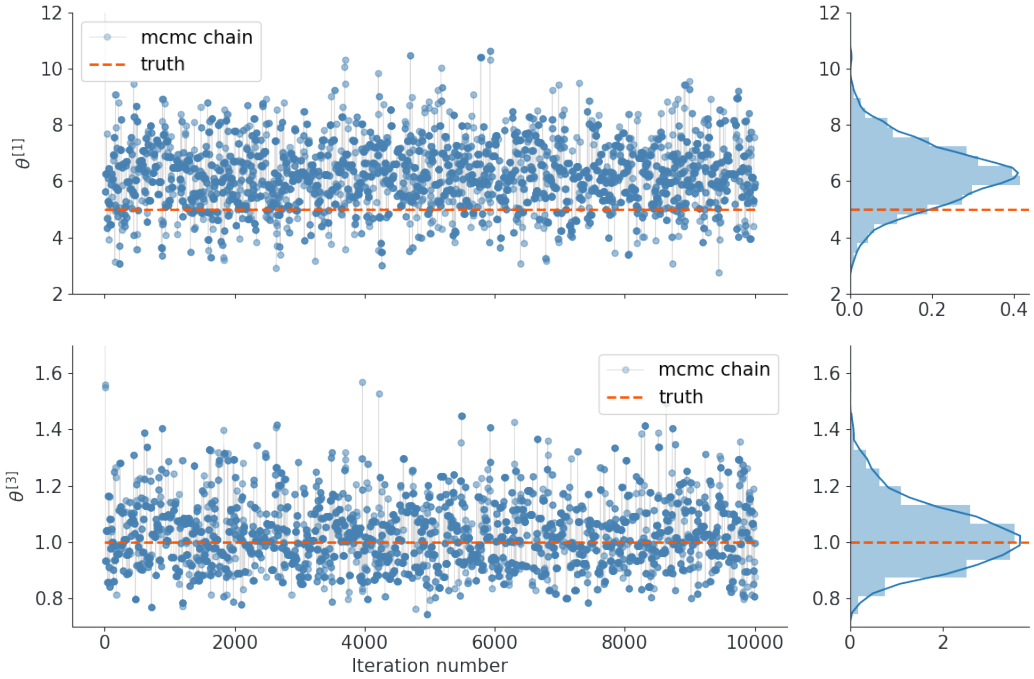


Figure 7.12: Traceplots from inference for the parameters $(\theta^{[1]}, \theta^{[3]})$ for the Langevin diffusion from eq. (7.28). The chain was initialised at $(\theta^{[1](0)}, \theta^{[3](0)}) = (20, 4)$.

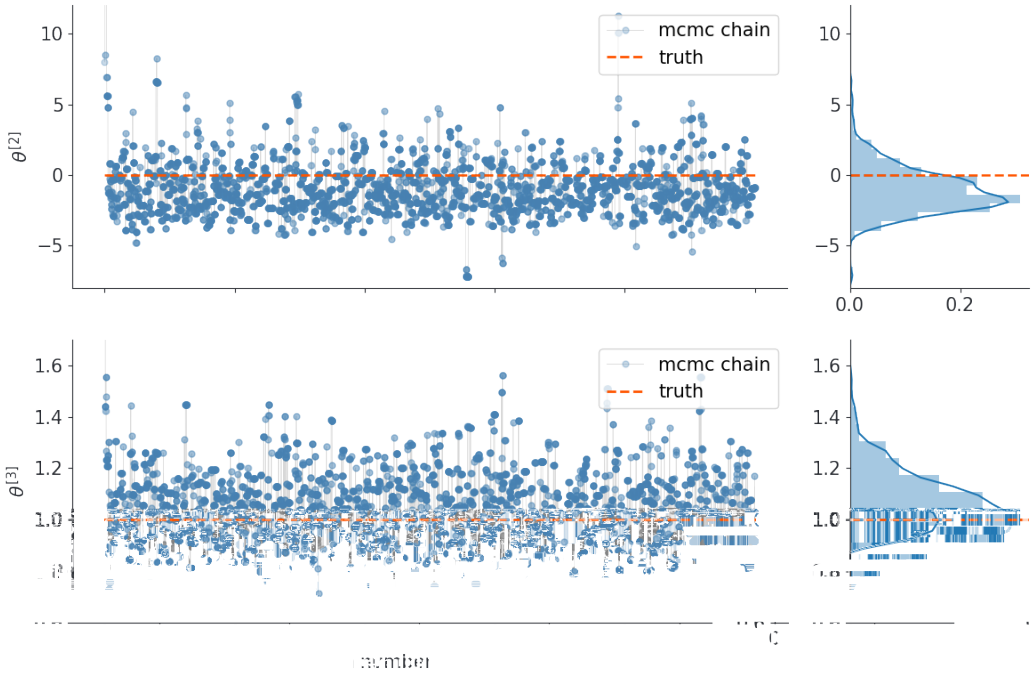


Figure 7.13: Traceplots from inference for the parameters $\theta^{[2:3]}$ for the Langevin diffusion from eq. (7.28). The chain was initialised at $\theta^{[2:3](0)} = (5, 4)$.

8

Conclusion

The central topic of this thesis has been the problem of simulating conditioned diffusions on a personal computer. Considered types of conditioning took various forms and included the terminal point of the process, a sequence of its partial and noisy observations, the first passage time to some fixed threshold or time-dependent barrier, as well as other, composite observational settings combining the previous ones. Throughout, I motivated the relevance of the solutions to this problem by using them in applications to Bayesian inference for diffusion processes—some of those applications culminated in novel inference results that could not have been obtained with alternative methods.

Two existing simulation algorithms have been instrumental to the developments of this thesis: rejection sampling on a path space due to Beskos and Roberts (2005) and guided proposals due to Schauer et al. (2017). They differ in their strengths and weaknesses, but in their own ways they are both at the forefront of relevance to the problem of simulating conditioned diffusions. The first algorithm distinguishes itself from any other, competing method in being devoid of any discretisation errors. Additionally, if implemented as a rejection sampler, it outputs independent draws of trajectories from the diffusion law of interest (revealed at a skeleton of points that, post-acceptance, can be filled-in at any other, arbitrary time-points). On the other hand, to apply the latter algorithm—guided proposals—the underlying diffusion needs to satisfy assumptions that often in practice are easy to satisfy. The diffusion matrix need not be invertible, Lamperti transformation need not exist, the drift need not be of a potential form and the process need not be uniformly elliptic (all of which are the pre-requisites for, say, rejection sampling on a path space) and yet, the algorithm may still apply (so long as other, weaker, technical conditions are satisfied). My aim in this thesis was to provide further extensions to those two methodologies.

In chapter 5 I examined the blocking technique in the context of simulating diffusion bridges. The methodology transforms the exact rejection sampler on a path space into a Gibbs sampler on a path space and induces a non-trivial trade-off of costs. I showed that the resulting algorithm, which is still devoid of any

discretisation errors, allows to reduce the scaling of the computational cost of the original algorithm from the exponential with the duration of the bridges down to being slightly in excess of cubic.

In chapter 6 I re-formulated core computational routines of guided proposals, which rendered the algorithm more efficient, as well as made it easier to write a generic implementation of it. To this end, I introduced a new set of backward ordinary differential equations, with solutions whose dimension is independent from the size of the dataset and which make it possible to compute all non-trivial terms required for a full implementation of guided proposals.

Finally, in chapter 7 I introduced a comprehensive treatment of the problem of simulating diffusions conditioned on their first passage times to some threshold (as well as related events). This allowed me to derive multiple Bayesian algorithms suitable for inference from first passage time observations—a problem relevant to a number of sciences. I focused primarily on the applications to neuroscience and I demonstrated that the introduced methodology can be used to efficiently find numerical solutions to first passage time inference problems for leaky integrate-and-fire models for a much broader range of the underlying diffusion processes than currently studied in the literature. Additionally, I introduced algorithms suitable for hypoelliptic diffusions conditioned on (composite) first passage time observations and I showed that the celebrated FitzHugh-Nagumo model, often employed in neuroscience, can be tackled with this methodology.

Going forward, there are a number of interesting directions that can be further developed. For blocking, further progress on the theory is needed so as to avoid the problem of relying on conjecture 5.2.1 and extend the results to inference. Additionally, one could try to come up with an adaptive scheme of knot placement so as to make maximal use of the available knots. Regarding the latter two chapters, guided proposals are steadily becoming a very flexible and efficient tool for simulating conditioned diffusions. With the results from chapter 6 I contributed to the development of the `Julia` package `BridgeSDEInference.jl`, which is a general implementation of the inference algorithm based on guided proposals that makes tackling any new inference problem straightforward to code. Integrating automated differentiation techniques and employing gradient-based inference methods

would allow to substantially speed-up the inference procedure and tackle higher dimensional problem with greater efficiency. To further improve the effectiveness of dealing with high-dimensional problems that exhibit elements of sparsity it would be invaluable to study numerical methods for sparse systems and develop appropriate modifications to the underlying algorithm. Space-discretisation of stochastic partial differential equations constitute an immensely broad and important group of process of this type. To tackle these with a satisfactory level of rigour, appropriate mathematical theory justifying use of guided proposals would have needed to be developed. Finally, there is a range of applications from physics, population dynamics and neuroscience for which the results on inference from first passage time observations could be applied to.

Reference list with all assumptions

Assumption A1. *Stochastic differential equation* (1.1) *admits a unique, weak, non-explosive solution.*

Condition C1. *Coefficients b and σ are locally Lipschitz continuous and grow at most linearly at infinity. I.e. for any compact set K there exists a constant c_K, such that for all $x, y \in K$:*

$$||\sigma(x)-\sigma(y)|| \leq c_K||x-y||, \quad ||b(x)-b(y)|| \leq c_K||x-y||,$$

and there exists a constant c, such that for all $x, y \in \mathbb{R}^d$:

$$||\sigma(x)|| \leq c(1+||x||), \quad ||b(x)|| \leq c(1+||x||).$$

Assumption A2. *Stochastic differential equation* (1.1) *admits a unique, strong, non-explosive solution.*

Assumption A3. *For all $x, y \in \mathbb{R}^d$ there exists an $\epsilon > 0$ such that $y^T\Gamma(x)y \geq \epsilon||y||^2$.*

Assumption A4. *Solution to stochastic differential equation* (1.1) *admits a smooth density.*

Assumption A5 (Novikov's condition)**.** $\mathbb{E}_\mu\left[\exp\left\{\frac{1}{2}\int_0^T [u^T u](X_t)\,\mathrm{d}t\right\}\right] < \infty$, *where expectation $\mathbb{E}_\mu$ is taken with respect to measure $\mathbb{P}_\mu$.*

Condition C2. *It is possible to obtain draws from $\mathbb{P}_\mu(\cdot|\mathcal{Z})$.*

Condition C3. *There exists a constant M^* satisfying* (2.5)*.*

Condition C4. *The Radon-Nikodým derivative $\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\mu}(X|\mathcal{Z})$ exists and is tractable.*

Assumption A6 (Lamperti transformation)**.** *σ is invertible and there exists a function $\eta : \mathbb{R}^d \to \mathbb{R}^d$, such that: $(\nabla_x \eta(x))^T = \sigma^{-1}(x)$.*

Assumption A7 (Potential form)**.** *The drift of a Lamperti transformed diffusion is of the potential form: i.e. there exists (a potential function) $A : \mathbb{R}^d \to \mathbb{R}$, such that $\nabla_x A(x) = \alpha(x)$.*

Assumption A8. *The drift of a Lamperti transformed diffusion is continuously differentiable:* $\alpha \in \mathcal{C}^1(\mathbb{R}^d; [0, T])$.

Assumption A9. $\mathcal{Z}$ *admits densities under the proposal* $\mathbb{P}_\mu$ *and the target* $\mathbb{P}_b$ *laws with respect to some dominating measure* $\varrho(\mathrm{d}\mathbb{z})$.

Assumption A10. *There exists a constant* $l_* > \infty$, *s.t.* $l_* \leq \inf\{\varphi(x);\, x \in \mathbb{R}^d\}$.

Assumption A11. *Function* $\upsilon(y) := \exp\{-||y-z||^2/2T + A(y)\}$ *is integrable for any* $z \in \mathbb{R}^d$.

Condition C5. $q(x, x^\circ) > 0$ *for all* $x, x^\circ \in supp(f)$.

Condition C6. *(Roberts and Tweedie, 1996)* f *is bounded on compact sets and there exist* $\epsilon > 0$ *and* $\delta > 0$, *for which* $q(x, x^\circ) > \epsilon$, *whenever* $d(x, x^\circ) < \delta$, *for some distance metric* d.

Assumption A12. *There exists a constant* $l^* > \infty$, *s.t.* $l^* \geq \sup\{\phi(x);\, x \in \mathbb{R}^d\}$.

Assumption A13. *The density of the speed measure*

$$m(x) := \frac{1}{\sigma^2(x)} \exp\left\{2 \int_0^x \frac{b(u)}{\sigma^2(u)}\, \mathrm{d}u\right\}, \quad x \in \mathbb{R},$$

of diffusion X *solving* (1.1) *is finite:* $\int_{\mathbb{R}} m(x)\, \mathrm{d}x < \infty$. *I.e. diffusion* X *is ergodic.*

Condition C7. *Volatility coefficients of the auxiliary and the target diffusions match at the end-point:* $\widetilde{\sigma}_T = \sigma(X_T)$.

Assumption A14. *(Bierkens et al., 2018, Assumption 2.7) There exists an invertible* $m \times m$ *diagonal matrix-valued function* $S(t)$ *(where* m *comes from the* $m \times d$ *observational operators* L_i*), which is measurable on [0,T), a* $t_0 < T$, $\gamma \in (0, 1]$ *and positive constants* $\underline{c}, \overline{c}, c_1, c_2$ *and* c_3 *such that for all* $t \in [t_0, T)$

$$\begin{aligned}
\underline{c}(T-t)^{-1} \leq \lambda_{\min}(M_S(t)) \leq \lambda_{\max}(M_S(t)) \leq \overline{c}(T-t)^{-1}, \\
\left\| [L_S(\widetilde{b} - b)](t, x) \right\| \leq c_1, \\
\mathrm{tr}\left([L_S \Gamma L_S^T](t, x)\right) \leq c_2, \\
\left\| [L_S(\widetilde{\Gamma} - \Gamma) L_S^T](t, x) \right\| \leq c_3 (T-t)^\gamma,
\end{aligned}$$

where $L_S(t) := [S\widetilde{L}](t)$, $M_S(t) := [S^{-1}\widetilde{M}S^{-1}](t)$, *and the pair* $\widetilde{L}(t)$, $\widetilde{M}(t)$ *are defined in eq.* (6.4) *and eq.* (6.7) *respectively.*

Condition C8. *Drift coefficients of the auxiliary and the target diffusions match at the end-point:* $\widetilde{b}_T = b(X_T)$.

Assumption A15. *The target law* $\mathbb{P}_b$ *is such that* $\mathcal{G}$ *is a Gaussian process.*

Assumption A16. *For* $t \in [0, T)$ $\widetilde{M}_t^\dagger$ *is invertible.*

Assumption A17. *The null-space of* $\widetilde{L}_t$, $t \in [0, T]$ *is equal to* $\{0\}$.

List of symbols

$\mathcal{C}(A;B)$	Space of continuous functions from $A \to B$
CBUS	Chequerboard blocking updating scheme
$\mathscr{D}$	Observations
$\mathscr{F}^*$	σ-algebra of the non-centrally parametrised probability space
$\mathscr{F}$	σ-algebra on $\mathscr{X}$
$\mathscr{F}_t$	Filtration
$\mathscr{G}$	Projection of the diffusion path on a discrete grid of time-points
Γ	Diffusion coefficient ($= \sigma\sigma^T$)
h	'h' function of the Doob's h-transform
$\widetilde{h}$	'h' function of the Doob's h-transform for the auxiliary diffusion
$\mathscr{H}$	Hypercube
$\mathscr{I}$	Sampled layer index
K	Total number of observations in $\mathscr{D}$
$\mathscr{K}$	Stochastic knots used in chapter 5
$\Bbbk$	Total number of transition kernels
$\varkappa$	Total number of points at which Poisson point process Φ is sampled
$\mathscr{L}$	Infinitesimal generator
$\mathbb{L}$	Law of the Poisson point process
l_*	Lower bound on the φ function inside the Radon-Nikodým derivative
l^*	Upper bound on the ϕ function inside the Radon-Nikodým derivative
L^*	Threshold level for first passage times
L_*	Renewal level
LBUS	Lexicographic blocking updating scheme
M^*	Upper bound on the ratio of the densities for a rejection sampler
$\mathbb{P}_b$	Target diffusion law (induced by SDE with drift b and volatility σ)
$\mathbb{P}_{b^\circ}$	Proposal diffusion law for guided proposals
$\mathbb{P}_\alpha$	Lamperti-transformed target diffusion law (induced by SDE with drift α and unit volatility coefficient)
$\mathbb{P}_0$	Wiener law
$\widetilde{\mathbb{P}}$	The law of the auxiliary diffusion
$\mathbb{Q}$	Diffusion law
$\mathbb{Q}^*$	Law of the non-centrally parametrised variable
RBUS	Random blocking updating scheme
$\mathscr{S}$	Surrogate random variable
τ^*	First passage time to a fixed threshold
τ_*	Renewal time
Υ	Layered information
$\mathscr{X}$	State-space
$\widetilde{X}$	Auxiliary diffusion
Φ	Poisson point process

χ_j	Time-positions of the sampled Poisson point process Φ
ψ_j	State-positions of the sampled Poisson point process Φ
Ψ_θ	Function transforming non-centrally parametrised variable to centrally parametrised one
$\mathbb{W}^*$	Law of Brownian bridges
$\mathbb{W}^*[T]$	Law of a 0–0 Brownian bridge defined on $[0, T]$
$\mathbb{W}$	Law of Brownian motion
$\mathbb{W}[T]$	Law of a standard Brownian motion defined on $[0, T]$
$\mathcal{Z}$	Conditioned-on random variable
$\widehat{\mathcal{Z}}$	Temporarily defined auxiliary conditioned-on random variable
$\mathbb{Z}$	Law of a biased Brownian motion
Ω^*	Sample space of the non-centrally parametrised probability space

Bibliography

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica 70*(1), 223–262.

Aït-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics 36*(2), 906–937.

Amit, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *Journal of Multivariate Analysis 38*(1), 82–99.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725.

Bachar, M., J. J. Batzel, and S. Ditlevsen (2012). *Stochastic biomathematical models: with applications to neuronal modeling*, Volume 2058. Springer.

Banon, G. (1978). Nonparametric identification for diffusion processes. *SIAM Journal on Control and Optimization 16*(3), 380–395.

Beskos, A., O. Papaspiliopoulos, and G. Roberts (2009). Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics 37*(1), 223–245.

Beskos, A., O. Papaspiliopoulos, and G. O. Roberts (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli 12*(6), 1077–1098.

Beskos, A., O. Papaspiliopoulos, and G. O. Roberts (2008). A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability 10*(1), 85–104.

Beskos, A., O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(3), 333–382.

Beskos, A., S. Peluchetti, and G. O. Roberts (2012). ε-Strong simulation of the Brownian path. *Bernoulli 18*(4), 1223–1248.

Beskos, A. and G. O. Roberts (2005). Exact simulation of diffusions. *The Annals of Applied Probability 15*(4), 2422–2444.

Beskos, A., G. O. Roberts, A. Stuart, and J. Voss (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics 8*(03), 319–350.

Beskos, A. and A. Stuart (2009). MCMC methods for sampling function space. In *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, Editors Rolf Jeltsch and Gerhard Wanner*, pp. 337–364.

Bibby, B. M., M. Jacobsen, and M. Sørensen (2010). Estimating functions for discretely sampled diffusion-type models. In *Handbook of financial econometrics: Tools and Techniques*, pp. 203–268. North-Holland.

Bierkens, J., F. van der Meulen, and M. Schauer (2018). Simulation of elliptic and hypo-elliptic conditional diffusions.

Bladt, M., S. Finch, and M. Sørensen (2016). Simulation of multivariate diffusion bridges. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(2), 343–369.

Bladt, M. and M. Sørensen (2014). Simple simulation of diffusion bridges with application to likelihood inference for diffusions. *Bernoulli 20*(2), 645–675.

Blanchet, J. and F. Zhang (2017). Exact Simulation for Multivariate Itô Diffusions.

Borodin, A. N. and P. Salminen (2002). *Handbook of Brownian motion: Facts and Formulae*. Birkhäuser.

Boys, R. J., D. J. Wilkinson, and T. B. L. Kirkwood (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing 18*(2), 125–135.

Brunel, N. and M. C. Van Rossum (2007). Lapicque's 1907 paper: from frogs to integrate-and-fire. *Biological cybernetics 97*(5-6), 337–339.

Bucy, R. S. and P. D. Joseph (2005). *Filtering for stochastic processes with applications to guidance*, Volume 326. American Mathematical Soc.

Casella, B. and G. O. Roberts (2011). Exact simulation of jump-diffusion processes with Monte Carlo applications. *Methodology and Computing in Applied Probability 13*(3), 449–473.

Cérou, F., P. Del Moral, T. Furon, and A. Guyader (2012). Sequential monte carlo for rare event estimation. *Statistics and computing 22*(3), 795–808.

Chen, N. and Z. Huang (2013). Localization and exact simulation of Brownian motion-driven stochastic differential equations. *Mathematics of Operations Research 38*(3), 591–616.

Chib, S., M. K. Pitt, and N. Shephard (2004). Likelihood based inference for diffusion driven models.

Clark, J. M. C. (1990). The simulation of pinned diffusions. In *29th IEEE Conference on Decision and Control*, pp. 1418–1420. IEEE.

Comte, F., V. Genon-Catalot, and Y. Rozenholc (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli 13*(2), 514–543.

Cotter, S. L., G. O. Roberts, A. M. Stuart, D. White, et al. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science 28*(3), 424–446.

Dacunha-Castelle, D. and D. Florens-Zmirou (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics: An International Journal of Probability and Stochastic Processes 19*(4), 263–284.

Dalalyan, A. S. and Y. A. Kutoyants (2002). Asymptotically efficient trend coefficient estimation for ergodic diffusion. *Mathematical Methods of Statistics 11*(4), 402–427.

Dayan, P. and L. F. Abbott (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press, Cambridge.

Delyon, B. and Y. Hu (2006, November). Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Processes and their Applications 116*(11), 1660–1675.

Devroye, L. (2006). Nonuniform random variate generation. *Handbooks in operations research and management science 13*, 83–121.

Di Nunno, G., B. K. Øksendal, and F. Proske (2009). *Malliavin calculus for Lévy processes with applications to finance*, Volume 2. Springer.

Diaconis, P., S. Holmes, and R. M. Neal (2000). Analysis of a nonreversible Markov chain sampler. *Annals of Applied Probability 10*(3), 726–752.

Ditlevsen, S. and O. Ditlevsen (2008). Parameter estimation from observations of first-passage times of the Ornstein-Uhlenbeck process and the Feller process. *Probabilistic Engineering Mechanics 23*(2-3), 170–179.

Ditlevsen, S. and P. Lansky (2005). Estimation of the input parameters in the Ornstein-Uhlenbeck neuronal model. *Physical review E 71*(1), 011907.

Ditlevsen, S. and P. Lansky (2006). Estimation of the input parameters in the Feller neuronal model. *Physical Review E 73*(6), 061910.

Ditlevsen, S. and P. Lansky (2007). Parameters of stochastic diffusion processes estimated from observations of first-hitting times: application to the leaky integrate-and-fire neuronal model. *Physical Review E 76*(4), 041906.

Ditlevsen, S. and A. Samson (2019). Hypoelliptic diffusions: filtering and inference from complete and partial observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(2), 361–384.

Doucet, A. and A. M. Johansen (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering 12*(656-704), 3.

Durham, G. B. and A. R. Gallant (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics 20*(3), 297–338.

Fearnhead, P., V. Giagos, and C. Sherlock (2014). Inference for reaction networks using the linear noise approximation. *Biometrics 70*(2), 457–466.

Fearnhead, P., K. Łatuszyński, G. O. Roberts, and G. Sermaidis (2017). Continuous-time importance sampling: Monte Carlo methods which avoid time-discretisation error.

Fearnhead, P., O. Papaspiliopoulos, and G. O. Roberts (2008). Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 755–777.

Fearnhead, P., O. Papaspiliopoulos, G. O. Roberts, and A. Stuart (2010). Random-weight particle filtering of continuous time processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 497–512.

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal 1*(6), 445–466.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics: A Journal of Theoretical and Applied Statistics 20*(4), 547–557.

Fortet, R. (1943). Les fonctions aléatoires du type de Markoff associées à certaines équations linéaires aux dérivées partielles du type parabolique. *J. Math. Pures. Appl. 22*(9), 177–243.

Fournié, E., J. M. Lasry, J. Lebuchoux, and P.-L. Lions (2001). Applications of Malliavin calculus to Monte Carlo methods in finance. II. *Finance and Stochastics 5*(2), 201–236.

Fuchs, C. (2013). *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Science & Business Media.

Geman, S. and D. Geman (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in computer vision*, pp. 564–584. Morgan Kaufmann.

Gobet, E., M. Hoffmann, and M. Reiß (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics 32*(5), 2223–2253.

Golightly, A. and D. J. Wilkinson (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology 13*(3), 838–851.

Golightly, A. and D. J. Wilkinson (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis 52*(3), 1674–1693.

Gonçalves, F. B., K. Łatuszyński, and G. O. Roberts (2017). Barker's algorithm for Bayesian inference with intractable likelihoods. *Brazilian Journal of Probability and Statistics 31*(4), 732–745.

Gugushvili, S., F. van der Meulen, M. Schauer, and P. Spreij (2018). Nonparametric Bayesian volatility learning under microstructure noise.

Hairer, M., A. M. Stuart, and J. Voss (2007). Analysis of SPDEs arising in path sampling part II: The nonlinear case. *The Annals of Applied Probability 17*(5/6), 1657–1706.

Hairer, M., A. M. Stuart, and J. Voss (2009). Sampling conditioned diffusions. *Trends in stochastic analysis 353*, 159–186.

Hairer, M., A. M. Stuart, J. Voss, and P. Wiberg (2005). Analysis of SPDEs arising in path sampling. part I: The Gaussian case. *Communications in Mathematical Sciences 3*(4), 587–603.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies 6*(2), 327–343.

Hodgkin, A. L. and A. F. Huxley (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology 117*(4), 500–544.

Hörmander, L. (1967). Hypoelliptic second order differential equations. *Acta Mathematica 119*(1), 147–171.

Huber, M. L. (2016). *Perfect simulation*. Chapman and Hall/CRC.

Imhof, J.-P. (1984). Density factorizations for Brownian motion, meander and the three-dimensional Bessel process, and applications. *Journal of Applied Probability 21*(3), 500–510.

Imkeller, P. and A. H. Monahan (2002). Conceptual stochastic climate models. *Stochastics and Dynamics 02*(3), 311–326.

Iolov, A., S. Ditlevsen, and A. Longtin (2017). Optimal design for estimation in diffusion processes from first hitting times. *SIAM/ASA Journal on Uncertainty Quantification 5*(1), 88–110.

Jacobsen, M. (1991). *Homogeneous Gaussian Diffusions in Finite Dimension*. Institute of Mathematical Statistics University of Copenhagen.

Jazwinski, A. H. (2007). *Stochastic processes and filtering theory*. Courier Corporation.

Jenkins, P. A. and D. Spano (2017). Exact simulation of the Wright–Fisher diffusion. *The Annals of Applied Probability 27*(3), 1478–1509.

Jensen, A. C., S. Ditlevsen, M. Kessler, and O. Papaspiliopoulos (2012). Markov chain monte carlo approach to parameter estimation in the fitzhugh-nagumo model. *Physical Review E 86*(4), 041114.

Kahn, H. and T. E. Harris (1951). Estimation of particle transmission by random sampling.

Kalogeropoulos, K. (2007). Likelihood-based inference for a class of multivariate diffusions with unobserved paths. *Journal of Statistical Planning and Inference 137*(10), 3092–3102.

Kalogeropoulos, K., G. O. Roberts, P. Dellaportas, et al. (2010). Inference for stochastic volatility models using time change transformations. *The Annals of Statistics 38*(2), 784–807.

Karatzas, I. and S. E. Shreve (1998a). *Brownian motion*. Springer.

Karatzas, I. and S. E. Shreve (1998b). *Methods of mathematical finance*, Volume 39. Springer.

Kelly, B. C., J. Bechtold, and A. Siemiginowska (2009). Are the variations in quasar optical flux driven by thermal fluctuations? *The Astrophysical Journal 698*(1), 895.

Kessler, M., A. Lindner, and M. Sørensen (2012). *Statistical methods for stochastic differential equations*. Chapman and Hall/CRC Press, NY.

Kloeden, P. E. and E. Platen (2013). *Numerical Solution of Stochastic Differential Equations*, Volume 23. Springer Science & Business Media.

Kulkarni, D., D. Schmidt, and S. K. Tsui (1999). Eigenvalues of tridiagonal pseudo-Toeplitz matrices. *Linear Algebra and its Applications 297*(1-3), 63–80.

Lansky, P. and S. Ditlevsen (2008). A review of the methods for signal estimation in stochastic diffusion leaky integrate-and-fire neuronal models. *Biological cybernetics 99*(4-5), 253.

Lauritzen, S. L. (1996). *Graphical models*, Volume 17. Clarendon Press.

Le Gall, J.-F. (2016). *Brownian motion, martingales, and stochastic calculus*, Volume 274. Springer.

Leon, J. R., A. Samson, et al. (2018). Hypoelliptic stochastic FitzHugh-Nagumo neuronal model: Mixing, up-crossing and estimation of the spike rate. *The Annals of Applied Probability 28*(4), 2243–2274.

Lindström, E. (2012). A regularized bridge sampler for sparsely sampled diffusions. *Statistics and Computing 22*(2), 615–623.

Lipster, R. S. and A. N. Shiryaev (2013). *Statistics of Random Processes I: General Theory*, Volume 5. Springer Science & Business Media.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the atmospheric sciences 20*(2), 130–141.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Meyn, S. P. and R. L. Tweedie (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Mullowney, P. and S. Iyengar (2008). Parameter estimation for a leaky integrate-and-fire neuronal model from ISI data. *Journal of Computational Neuroscience 24*(2), 179–194.

Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE 50*(10), 2061–2070.

Nicolau, J. (2002). A new technique for simulating the likelihood of stochastic differential equations. *The Econometrics Journal 5*(1), 91–103.

Øksendal, B. (2013). *Stochastic differential equations: an introduction with applciations*. Springer Sience & Business Media.

Papaspiliopoulos, O. (2003). *Non-centered parameterisations for data augmentation and hierarchical models*. Ph. D. thesis, Lancaster University.

Papaspiliopoulos, O., Y. Pokern, G. O. Roberts, and A. M. Stuart (2012). Nonparametric estimation of diffusions: a differential equations approach. *Biometrika 99*(3), 511–531.

Papaspiliopoulos, O. and G. Roberts (2012). Importance sampling techniques for estimation of diffusion models. *Statistical methods for stochastic differential equations 124*, 311–340.

Papaspiliopoulos, O., G. O. Roberts, and M. Sköld (2003). Non-centered parameterisations for hierarchical models and data augmentation. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, Volume 307. Oxford University Press, USA.

Papaspiliopoulos, O., G. O. Roberts, and O. Stramer (2013). Data augmentation for diffusions. *Journal of Computational and Graphical Statistics 22*(3), 665–688.

Papaspiliopoulos, O., G. O. Roberts, and K. B. Taylor (2016). Exact sampling of diffusions with a discontinuity in the drift. *Advances in Applied Probability 48*(A), 249–259.

Papaspiliopoulos, O., G. O. Roberts, and R. L. Tweedie (in prep). The methodology, ergodicity and optimisation of Markov chain Monte Carlo. (In preparation).

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian journal of statistics 22*(1), 55–71.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika 60*(3), 607–612.

Pitman, J. and M. Yor (1982). A decomposition of Bessel bridges. *Probability Theory and Related Fields 59*(4), 425–457.

Pokern, Y., A. M. Stuart, and J. H. van Zanten (2013). Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stochastic Processes and their Applications 123*(2), 603–628.

Pollock, M. (2013, 9). *Some Monte Carlo methods for jump diffusions*. Ph. D. thesis, University of Warwick.

Pollock, M., A. M. Johansen, G. O. Roberts, et al. (2016). On the exact and ε-strong simulation of (jump) diffusions. *Bernoulli 22*(2), 794–856.

Poulsen, R. (1999). *Approximate maximum likelihood estimation of discretely observed diffusion processes*. CAF, Centre for Analytical Finance, University of Aarhus.

Ricciardi, L. M., A. D. Crescenzo, V. Giorno, and A. Nobile (1999). An outline of theoretical and algorithmic approaches to first passage time problems with applications to biological modeling. *Mathematica Japonica 50*, 247–322.

Ricciardi, L. M. and S. Sato (1988). First-passage-time density and moments of the Ornstein-Uhlenbeck process. *Journal of Applied Probability 25*(1), 43–57.

Robert, C. and G. Casella (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Roberts, G. and J. Rosenthal (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability 2*, 13–25.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(2), 291–317.

Roberts, G. O. and O. Stramer (2001). On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika 88*(3), 603–621.

Roberts, G. O. and R. L. Tweedie (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika 83*(1), 95–110.

Rogers, L. C. G. and D. Williams (2000a). *Diffusions, Markov processes, and martingales: Volume 1, foundations*, Volume 1. Cambridge university press.

Rogers, L. C. G. and D. Williams (2000b). *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, Volume 2. Cambridge university press.

Schauer, M., F. van der Meulen, and H. van Zanten (2017). Guided proposals for simulating multi-dimensional diffusion bridges. *Bernoulli 23*(4A), 2917–2950.

Schlick, T. (2010). *Molecular modeling and simulation: an interdisciplinary guide*, Volume 21. Springer Science & Business Media.

Sermaidis, G., O. Papaspiliopoulos, G. O. Roberts, A. Beskos, and P. Fearnhead (2013, June). Markov chain Monte Carlo for exact inference for diffusions. *Scandinavian Journal of Statistics 40*(2), 294–321.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*(3), 653–667.

Shoji, I. (1998). Approximation of continuous time stochastic processes by a local linearization method. *Mathematics of Computation of the American Mathematical Society 67*(221), 287–298.

Shoji, I. and T. Ozaki (1998a). Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications 16*(4), 733–752.

Shoji, I. and T. Ozaki (1998b). A statistical method of estimation and simulation for systems of stochastic differential equations. *Biometrika 85*(1), 240–243.

Siegert, A. J. (1951). On the first passage time probability problem. *Physical Review 81*(4), 617.

Smith, G. D. (1985). *Numerical solution of partial differential equations: finite difference methods*. Oxford University Press.

Sobczyk, K. (2013). *Stochastic differential equations: with applications to physics and engineering*, Volume 40. Springer Science & Business Media.

Sparrow, C. (2012). *The Lorenz equations: bifurcations, chaos, and strange attractors*, Volume 41. Springer Science & Business Media.

Spokoiny, V. G. (2000). Adaptive drift estimation for nonparametric diffusion model. *The Annals of Statistics 28*(3), 815–836.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *The Journal of Finance 52*(5), 1973–2002.

Stramer, O. and M. Bognar (2011). Bayesian inference for irreducible diffusion processes using the pseudo-marginal approach. *Bayesian Analysis 6*(2), 231–258.

Stramer, O. and G. O. Roberts (2007). On bayesian analysis of nonlinear continuous-time autoregression models. *Journal of Time Series Analysis 28*(5), 744–762.

Stuart, A. M., J. Voss, and P. Wilberg (2004). Conditional path sampling of SDEs and the langevin MCMC method. *Communications in Mathematical Sciences 2*(4), 685–697.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics 22*(4), 1701–1728.

Tuan, P. D. (1981). Nonparametric estimation of the drift coefficient in the diffusion equation. *Series Statistics 12*(1), 61–73.

Tuckwell, H. C. (1988). *Introduction to Theoretical Neurobiology*, Volume 1 of *Cambridge Studies in Mathematical Biology*. Cambridge University Press.

van der Meulen, F. and M. Schauer (2017a). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics 11*(1), 2358–2396.

van der Meulen, F. and M. Schauer (2017b). Continuous-discrete smoothing of diffusions.

van der Meulen, F. and M. Schauer (2018). Bayesian estimation of incompletely observed diffusions. *Stochastics 90*(5), 641–662.

van der Meulen, F., M. Schauer, and J. van Waaij (2018). Adaptive nonparametric drift estimation for diffusion processes using faber–schauder expansions. *Statistical Inference for Stochastic Processes 21*(3), 603–628.

van der Meulen, F., M. Schauer, and H. van Zanten (2014). Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *Computational Statistics & Data Analysis 71*, 615–632.

van der Pol, B. and J. van der Mark (1928). LXXII. the heartbeat considered as a relaxation oscillation, and an electrical model of the heart. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6*(38), 763–775.

van Waaij, J. and H. van Zanten (2016). Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electronic Journal of Statistics 10*(1), 628–645.

van Zanten, H. (2001). Rates of convergence and asymptotic normality of kernel estimators for ergodic diffusion processes. *Journal of Nonparametric Statistics 13*(6), 833–850.

van Zanten, H. (2013). Nonparametric bayesian methods for one-dimensional diffusion models. *Mathematical Biosciences 243*(2), 215–222.

Vanden-Eijnden, E. and J. Weare (2012). Rare event simulation of small noise diffusions. *Communications on Pure and Applied Mathematics 65*(12), 1770–1803.

Verner, J. H. (1978). Explicit Runge–Kutta methods with estimates of the local truncation error. *SIAM Journal on Numerical Analysis 15*(4), 772–790.

von Neumann, J. (1963). Various techniques used in connection with random digits. *John von Neumann, Collected Works 5*, 768–770.

Wang, L. and K. Pötzelberger (1997). Boundary crossing probability for Brownian motion and general boundaries. *Journal of Applied Probability 34*(1), 54–65.

Whitaker, G. A., A. Golightly, R. J. Boys, and C. Sherlock (2017). Improved bridge constructs for stochastic differential equations. *Statistics and Computing 27*(4), 885–900.

Williams, D. (1970). Decomposing the brownian path. *Bulletin of the American Mathematical Society 76*(4), 871–873.

Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis 41*(2), 220–242.

Zhang, X., G. You, T. Chen, and J. Feng (2009). Maximum likelihood decoding of neuronal inputs from an interspike interval distribution. *Neural Computation 21*(11), 3079–3105.