SIMULATION OF DIFFUSION BRIDGES FOR STOCHASTIC DIFFERENTIAL EQUATIONS

MASTER'S THESIS IN MATHEMATICAL STATISTICS

BY

DANIEL DAMBERG

SUPERVISOR:
PROF. ERIK LINDSTRÖM

LUND
UNIVERSITY

FACULTY OF ENGINEERING
2017

Abstract

The simulation of diffusion bridges is a difficult statistical problem, and methods for this are attracting a great deal of attention. Analytically, the conditioning on a future point introduces a second term to the drift of the SDE that includes the transition density of the unconditioned process. However, transition densities of diffusion processes are often intractable, whereby other, often computationally intensive methods are necessary. A useful method is Monte Carlo simulation, the computational cost of which, however, is often prohibitively high, necessitating the use of some variance reduction techniques.

In this thesis, a novel method for simulating discrete time realisations of diffusion processes satisfying some stochastic differential equation (SDE), and conditioned on hitting some end-point is proposed. This new method is based on the partitioning of the process into one deterministic and one random part, where the deterministic part accounts for the drift of the process, leaving the residual, random process a Brownian bridge. The difficulty then lies in finding a good approximation of the deterministic process, and it is herein proposed that already accepted realisations are used to adaptively improve an estimate of the optimal deterministic process, which is used for simulation.

The proposed sampler uses the linear noise approximation (LNA) of the process as initial distribution. Taking the mean of previous realisations ensures convergence to the proper expected path, but may initially produce distorted approximations. During an initial burn-in period, the proposed sampler therefore uses recursively formulated regression based on a suitably chosen number of sine basis functions to approximate the remaining dynamics and forms the approximation of the expected path as the sum of the LNA and the approximation of the remainder. Choosing a suitably low number of basis functions accords the regressive approximation with a resilience towards the erratic behaviour of individual realisations but introduces bias. Therefore, after a burn-in period, the sampler switches to taking the mean, ensuring proper convergence.

The proposed sampler is tested on two different models: the Cox-Ingersoll-Ross model and the stochastic Lorenz model, and its performance compared to existing approaches, showing similar results for easy cases and a significant increase for difficult ones.

Keywords: Diffusion bridge, Markov chain Monte Carlo, Stochastic differential equation, Linear noise approximation

Acknowlegements

I should like to express my sincere gratitude to my supervisor, Professor Erik Lindström, to whom I shall always remain indebted, for his ideas, his advice and his unfailing encouragement.

Furthermore, I direct my thanks to Fabian, Lukas, Sebastian, Linus, Hampus, and Fredrik, without whose congenial company this process would have been a decidedly less enlivening experience.

Finally, I should like to thank my family for their never-ending support and encouragement, and for instilling in me an ever-present sense of curiosity.

Contents

Chapter 1

Introduction

1.1 Background

In many fields, the attempt to mathematically model various processes of interest leads to the formulation of (a system of) differential equations. A simple example is Newton's well known second law of motion:

$$m\frac{\mathrm{d}^2}{\mathrm{d}t^2}x(t) = F(x(t)).$$

However, this formulation, and others like it, is based upon a host of approximations, which can be both conceptual and mathematical. For example, Newton did not take the effects of relativity or quantum mechanics into account when he formulated his laws of motion, and experience has shown this to be justified. But empiricisms aside, these additional effects, albeit negligible at speeds far from the speed of light and in our macroscopic world, are there, and their absence from Newton's laws of motion, in some sense, constitute an approximation.

The addition of a random term may serve to account for certain approximations. An epoch-making example is Brownian motion. The force acting on a particle in a liquid is traditionally given by Stokes' law, but physicist Paul Langevin, in his seminal 1908 note 'Sur la théorie du mouvement brownien', remarks that the aforementioned value *»n'est qu'une moyenne»*, proposing the addition of a complementary force, positive or negative, to maintain the erratic movement that would otherwise have been stopped by viscous resistance.

There are plenty of examples of models that would be well served by the inclusion of a random term in the usual differential equation. However, the stringent mathematical definition of such a random differential equation is not simple, and it was not until the middle of the 20th century that stochastic differential equations were given a solid mathematical foundation.

One is interested in the stochastic processes, X, whose local dynamics may be approximated by a stochastic difference equation of the following form,

$$X(t+\delta) - X(t) = \mu(t, X(t))\Delta t + \sigma(t, X(t))G(t), \tag{1.1}$$

where μ and σ are deterministic functions, and $G(t)$ is a Gaussian disturbance term, independent of the history of the process until present time, t. The process is then called

a diffusion process with *drift* μ and *diffusion* σ. The pertinent differential equation is normally written

$$\begin{aligned} dX_t &= \mu(t, X(t))dt + \sigma(t, X(t))dW(t), \\ X_0 &\in \mathbb{R}^d. \end{aligned} \tag{1.2}$$

(Björk 2009)

A diffusion bridge is a process X^*, obeying the same stochastic differential equation as above, but conditioned to hit some point $v \in \mathbb{R}^d$ at a later time $T > 0$. Thus, it forms a bridge between some initial point u and the final point v.

For stochastic differential equation models like the one above, transition densities are often intractable. Only the simplest models, e.g. the Ornstein-Uhlenbeck process and the Cox-Ingersoll-Ross process, have transition densities that can be obtained on a manageable form, with most 'interesting' models requiring numerical methods in order to obtain their transition densities.

The transition densities are important for, e.g. parameter estimation and model selection, with likelihood functions being based upon them. Due to the aforementioned difficulty, the methods employed to solve these estimation problems are often computationally expensive, necessitating the use of variance reduction techniques.

As a specific example one can regard the replacement of missing data, often referred to as data imputation. Although the aforementioned process X is continuous, the data available, whether in finance or in other areas, is often sampled at—perhaps low-frequency—discrete time points. With transition densities, and thereby the likelihood for the data, probably being intractable, the estimation problem becomes quite difficult. It has been proposed that the segments between the data points be regarded as missing data. These gaps may then be filled by simulating diffusion bridges between the data points, and many methods have been proposed to efficiently augment data thus. The robustness of such methods can furthermore be of importance when using e.g. Gibbs sampling for parameter estimation, and they should be able to handle extreme cases (with regards to the parameters examined), which may occur if the examined parameters are not close to the true ones.

1.2 Objectives and limitations

The primary goal of this thesis is to find an improved method for simulating diffusion bridges for stochastic differential equations, i.e. processes conditioned to hit a specified end-point. It examines some, already existing methods for the this purpose, and, with these methods as foundation and keeping in mind the Metropolis-Hastings framework in which simulations are carried out, attempts to use adaptive Markov chain Monte Carlo to find the expected path of the process, in order to use this to create a better algorithm for simulating diffusion bridges.

The improved method is evaluated and compared with existing methods in terms of statistical efficiency, measured via their estimated acceptance ratios, and with regards to computational efficiency – although computational efficiency is only treated very briefly.

1.3 Disciplinary foundations

Simulating diffusion bridges with both statistical and computational efficiency is a challenging task, and a great deal has been written on the subject. Simulation proced-

ures often use an acceptance-rejection method, e.g. Metropolis-Hastings algorithms. Important texts in the subject include the following articles.

A simple simulation approach based upon the use Euler-Maruyama discretisation for proposal densities was proposed by Pedersen 1995.

Durham and Gallant 2002 propose what they call a 'Modified diffusion bridge', wherein they, in generating step $m + 1$ of the bridge, condition on the previous step m as well as the end-point. This approach, however, is less effective if the process exhibits non-linear dynamics.

Erik Lindström 2012 combines the Pedersen and the Durham and Gallant samplers to provide variance reduction for sparsely sampled data.

Whitaker et al. 2016 propose the partitioning of the process into two processes, one that accounts deterministically for the non-linear dynamics, and one residual process upon which the modified diffusion bridge method may be used.

1.4 Contribution and organisation

In this report, a new and improved algorithm for the simulation of diffusion bridges is presented. The new method is based on the residual bridge constructs of Whitaker et al. 2016, but introduces an adaptive updating of the deterministic process, taking it progressively closer to the true expected path of the conditioned process. Like in Whitaker et al. 2016, the residual, stochastic process—the dynamics of which are (hopefully) linear—is then simulated using the modified diffusion bridge construct. The target bridge is obtained by summing the residual bridge and the approximate, deterministic expected path.

Chapter 2 gives a broad, theoretical foundation, presenting the necessary mathematical and probabilistic concepts. Chapter 3 recounts important existing methods for simulating diffusion bridges in some detail. Next, the proposed, new sampler is introduced in Chapter 4, followed by a simulation study of the performance of the proposed sampler compared with the samplers of Whitaker et al. 2016 in Chapter 5. Concluding the report, a discussion is provided in Chapter 6.

Chapter 2

Theory

2.1 Markov chain Monte Carlo methods and Bayesian statistics

We begin with a brief exposition of some methods for the obtention of expectations of functions of random variables, which are often important quantities in a statistical analysis. For example, with an appropriate indicator function, one may obtain distribution functions. First, however, we take a brief look at Bayesian statistics.

2.1.1 Bayesian framework

This particular school of thought takes its name from the English 18th century statistician and theologian Thomas Bayes, whose famous theorem on conditional probabilities lies at the heart of the paradigm. A core concept of the Bayesian paradigm is its interpretation of probability, where, philosophical finesse aside, prior knowledge of the object of one's studies is taken into account.

More practically, one associates probability distributions with the parameters of the likelihood, effectively making them random variables. This distribution may then be used to assign relative probabilities to different regions of the parameter space in an effort to quantify one's knowledge about the parameters.

Phrased in a mathematical way, it is supposed that X has a distribution that is parameterised by θ and assign the *a priori* density $f(\theta)$ to θ before any observation has been made. This prior distribution may be chosen based on previous knowledge, purely personal beliefs, or in a manner such that its influence on the final inference is as limited as possible. Now, data are needed. The data yield a likelihood $L(\theta|x)$ that gives information about the parameters, whereby prior beliefs about the parameters can be updated. One's knowledge is updated with Bayes' theorem (hence the name),

$$f(\theta|x) \propto f(\theta)f(x|\theta) = f(\theta)L(\theta|x), \tag{2.1}$$

whence the *a posteriori* density of the parameters is obtained. This may then be used for inference about θ. This approach is fundamentally different from the frequentist approach, where parameters are considered to be unknown but fixed. Naturally, there is a proportionality constant, which has been omitted above, equalling $1/\int f(\theta)L(\theta|x)d\theta$. This constant is often difficult to compute—and for certain purposes not necessary—

and the numerical methods hereafter described are often used for estimation of this constant.

The posterior mode, denoted $\tilde{\theta}$, is a consistent estimator of the true parameter values θ^*, converging to a $N(\theta^*, I(\theta^*)^{-1})$ as the number of data points $n \to \infty$, where $I(\theta^*)$ is the Fisher information matrix. Here, we can also see that the data should overwhelm any prior in the limit, which is obviously wanted. (Givens and Hoeting 2012)

2.1.2 Importance sampling

Now, let us return to the expectations mentioned above, $\mu = \mathbb{E}[h(X)]$. With f denoting the density of X, the expectation μ can be approximated by a sample average:

$$\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{a.s.} \int h(x)f(x)dx = \mu, \tag{2.2}$$

where X_i are draws from f. The convergence follows from the strong law of large numbers. Assuming further that $h(X)^2$ has finite expectation under f, the variance of the estimator is $\sigma^2/n = \mathbb{E}^f[(h(x) - \mu)^2/n]$. An estimation of σ^2 is then given, in a similar fashion as above, as

$$\hat{\mathbb{V}}[\hat{\mu}_{MC}] = \frac{1}{n-1}\sum_{i=1}^{n}(h(X_i) - \hat{\mu}_{MC})^2. \tag{2.3}$$

When σ^2 exists, the central limit theorem implies that $\hat{\mu}_{MC}$ has an approximate normal distribution for large n. (Givens and Hoeting 2012)

The aforementioned method for simulating expectations (or really calculating integrals by means of simulation) is called the Monte Carlo method, presumably as a reference to the famous Monégasque casino.

There are some (potential) problems with the crude Monte Carlo method. For example, if the sought-after density f is difficult to sample from, or if the integrand h and the density f are dissimilar, and the lion's share of the draws from f land in an area where the integrand is minute, crude Monte Carlo simulation becomes computationally inefficient. In the first example, the difficulty is self-evident, and in the second, only very few draws will actually contribute to forming the sample mean, yielding a large variance.

A solution to the aforementioned problems is importance sampling, i.e. making interesting events occur more often than what would have been the case with the naïve method above. Thereby one obtains increased accuracy. Now, one has also distorted the sampling distribution, 'oversampling' parts of the density, so one has to introduce an importance weighting to correct for this distortion. Very large variance reductions can be obtained when the integrand and the distribution are very dissimilar.

In practice, importance sampling means the introduction of a second *importance sampling* density, $g(x)$, and the multiplication with a 'creative one', yielding

$$\mu = \mathbb{E}_f[h(X)] = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = \mathbb{E}_g\left[h(X)\frac{f(x)}{g(x)}\right]. \tag{2.4}$$

Hence, by similar arguments as for the crude Monte Carlo above, a Monte Carlo estimate of $\mathbb{E}[h(x)]$ is given by drawing i.i.d. samples from g and computing

$$\hat{\mu}^*_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w^*(X_i), \tag{2.5}$$

where $w^*(X_i) = f(X_i)/g(X_i)$ are unstandardised weights, called importance ratios. Equation (2.4) can also be written

$$\mu = \frac{\int h(x)f(x)dx}{\int f(x)dx} = \frac{\int h(x)[f(x)/g(x)]g(x)dx}{\int [f(x)/g(x)]g(x)dx}, \tag{2.6}$$

leading to the standardised estimator

$$\hat{\mu}_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w(X_i), \tag{2.7}$$

where $w(X_i) = w^*(X_i)/\sum_{j=1}^{n} w^*(X_j)$ are standardised weights. The first estimator is convenient when it is easy to sample from g and easy to evaluate f. The second, standardised estimator can be used even when f is only known up to a normalising constant, as is often the case when dealing with a posteriori densities under the Bayesian inferential paradigm.

For convergence to the desired quantity, it is necessary that the support of g covers all of the support of f. Furthermore, $f(x)/g(x)$ should be bounded, and g should have heavier tails than f, so as to avoid undue variability. The estimators both converge by the same argument as for crude Monte Carlo above. (Givens and Hoeting 2012)

It can be readily seen that the optimal proposal distribution g is proportional to the absolute value of the function of interest times the original density. This is unfortunately not something that can be directly used in practice—the desired quantity plays a part in the expression—but it may be significant in theoretical analyses.

Theorem 2.1. *The proposal density that minimises the variance of the importance sampler* $\hat{\mu}_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)[f(X_i)/g(X_i)]$ *is*

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(y)|f(y)dy}. \tag{2.8}$$

Proof. The variance of the importance sampler is given by

$$\mathbb{V}_g\left[h(X)\frac{f(x)}{g(x)}\right] = \mathbb{E}_g\left[h(X)^2\frac{f(x)^2}{g(x)^2}\right] - \mathbb{E}_g\left[h(X)\frac{f(x)}{g(x)}\right]^2. \tag{2.9}$$

Let g be any density which is positive whenever $h(x)f(x) \neq 0$, and denote the desired expectation by μ and the variance under proposal density g by σ_g^2. First, consider the squared mean,

$$\begin{aligned}\mu^2 + \sigma_{g^*}^2 = \mathbb{E}_{g^*}\left[h(X)^2\frac{f(x)^2}{g^*(x)^2}\right] = \int h(x)^2\frac{f(x)^2}{g^*(x)}dx = \int \frac{h(x)^2 f(x)^2}{\frac{|h(x)|f(x)}{\int |h(y)|f(y)dy}}dx \\ = \int |h(x)|f(x)dx\left(\int |h(y)|f(y)dy\right) = \mathbb{E}_f[|h(X)|]^2.\end{aligned} \tag{2.10}$$

Now introduce a new proposal density g, i.e. not the optimal one, and rewrite the expression accordingly. Then, by Jensen's inequality, we obtain that our optimal g^* is smaller than or equal to an arbitrary proposal density g,

$$\begin{aligned}\mu^2 + \sigma_{g^*}^2 = \mathbb{E}_f[|h(X)|]^2 = \mathbb{E}_g\left[|h(X)|\frac{f(X)}{g(X)}\right]^2 \\ \leq \mathbb{E}_g\left[|h(X)|^2\frac{f(X)^2}{g(X)^2}\right] = \mu^2 + \sigma_g^2,\end{aligned} \tag{2.11}$$

i.e. $\sigma^2_{g^*} \leq \sigma^2_g$, which is the desired result. □

This can be put into a more formal probabilistic framework (to be introduced in section 2.2), where, instead of looking at distributions as above, one regards the pertinent probability measures. Let the sample space Ω, the σ-algebra of subsets of Ω, $\mathscr{F}$ and the probability measure $\mathbb{P}$, i.e. the triple $(\Omega, \mathscr{F}, \mathbb{P})$, be a probability space. The desired expectations can then be written

$$\mu = \int h(x) d\mathbb{P}(x). \tag{2.12}$$

Importance sampling then constitutes the change of measure to the new probability measure $\mathbb{Q}$, absolutely continuous with respect to $\mathbb{P}$, and the importance weights $f(x)/g(x)$ is the Radon-Nikodym derivative $\frac{d\mathbb{P}}{d\mathbb{Q}}$. The estimator will then be

$$\hat{\mu}_{IS} = \frac{1}{n} \sum_{i=1}^{n} h(x) \frac{d\mathbb{P}}{d\mathbb{Q}}. \tag{2.13}$$

2.1.3 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm, introduced by Metropolis et al. 1953 and generalised by Hastings 1970, is a famous, if not the most famous, Markov chain Monte Carlo method. Markov chain Monte Carlo methods are used for generating draws from a distribution that approximates f, or, perhaps more properly, methods for generating samples from which the type of expectations discussed earlier, $\mathbb{E}[h(x)]$, can be calculated. It is an easily customisable class of methods that can handle a wide range of different—and difficult—problems. Herein lies their strength. The basic idea behind Markov chain Monte Carlo is to construct a Markov chain with stationary distribution f. (Givens and Hoeting 2012)

First, a fundamental definition. Some necessary definitions are found in Section 2.2.1.

Definition 2.2. (Markov chain) *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space with a non-decreasing family of σ-algebras $\{\mathscr{F}\}_{t\geq 0}$, such that $\mathscr{F}_0, \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}$. Then a stochastic sequence $X = (X_t, \mathscr{F}_t)$ is called a* Markov chain *with respect to $\mathbb{P}$ if*

$$\mathbb{P}[X_t \in B | \mathscr{F}_s] = \mathbb{P}[X_t \in B | X_s] \quad (\mathbb{P} - a.s.), \tag{2.14}$$

for all $t \geq s \geq 0$ and all $B \in \mathscr{B}(\mathbb{R})$.

That is to say, for any given 'present', the 'future' is independent of the 'past'. (Shiryayev and Boas 1984) Hence, we have the following result.

Theorem 2.3. *Let $\{X_t\}$ be a (discrete-time) Markov chain with initial distribution ξ. Then the joint density is given by*

$$f(x_0, x_1, \ldots, x_T) = \xi(x_0) \prod_{t=0}^{T-1} q(x_{t+1} | x_t), \tag{2.15}$$

for $t \geq 0$, where q is the transition density of the chain.

Proof. By the chain rule, the joint distribution can be written

$$P(X_0, \ldots, X_T) = P(X_T|x_0, \ldots, x_{T-1}) \cdots P(X_1|x_0)P(X_0), \tag{2.16}$$

which, by the Markov property, reduces to

$$P(X_0, \ldots, X_T) = P(X_T|x_{T-1})P(X_{T-1}|x_{T-2}) \cdots P(X_1|x_0)P(X_0). \tag{2.17}$$

The density of X_0 has, by definition, been set to $\xi(x_0)$. The result easily follows. □

Straightforward marginalisation yields the following corollary.

Corollary 2.4. (Chapman-Kolmogorov equation) *Let $\{X_t\}$ be a discrete-time Markov chain. Then*

$$f(X_t|x_0) = \int \cdots \int \left(\prod_{t=0}^{T-1} q(x_{t+1}|x_t) \right) dx_1 \cdots dx_{t-1}. \tag{2.18}$$

An irreducible and aperiodic Markov chain converges to a limiting stationary distribution, a fact that can be helpful in simulating from a specified distribution. If one manages to construct a Markov chain with the desired distribution as stationary distribution, a realisation from that chain will, for sufficiently large times t, approximately come from the desired distribution. The objective of Markov chain Monte Carlo is to construct such chains.

The Metropolis-Hastings algorithm begins with the initialisation of the chain, setting $X_0 = x_0$, drawn randomly from some initial distribution g. It is essential (obviously) that we have $f(x_0) > 0$.

In Metropolis-Hastings sampling, one samples from a proposal distribution, from which it is easier to generate samples, and then accepts or rejects the proposals based on information about the proposal and the target. A description of the algorithm can be found in Algorithm 2.1.

Algorithm 2.1 The Metropolis-Hastings algorithm. (Givens and Hoeting 2012)

1: Initialise x_0.
2: **Repeat** for all t:
3: Given x_{t-1}, generate a candidate X^* from a proposal distribution $g(\cdot|x_{t-1})$.
4: Compute the Metropolis-Hastings ratio $R(x_{t-1}, X^*)$, where

$$R(u, v) = \frac{f(v)g(u|v)}{f(u)g(v|u)}. \tag{2.19}$$

5: Set the next value, X_t, as

$$X_t = \begin{cases} X^*, & \text{with probability } \min\{R(x_t, X^*), 1\}, \\ x_{t-1}, & \text{otherwise.} \end{cases} \tag{2.20}$$

6: **End**

In some cases, effects of the choice of initial distribution lingers, deteriorating the performance of the algorithm. It may then be sensible to omit some early realisations, and delay one's analysis until these initial effects have receded. This first period,

the realisations of which are discarded, is called a burn-in period and is an important tool in Markov chain Monte Carlo. Once a satisfactory chain has been obtained, the expectation can be approximated as the average over realisations from the stationary distribution. (Givens and Hoeting 2012)

If one chooses the proposal distribution without dependence on earlier realisations, i.e. $g(x^*|x_t) = g(x^*)$, then one obtains an independence chain. The Metropolis-Hastings ratio reduces to

$$R(x_t, X^*) = \frac{f(X^*)g(x_t)}{f(x_t)g(X^*)}, \tag{2.21}$$

and the resulting chain is irreducible and aperiodic, that is, there is a stationary distribution to which the chain converges, if $g(x) > 0$ whenever $f(x) > 0$. Furthermore, g should be chosen such that it resembles f but still covers the tails of f.

2.1.4 Gibbs sampling

Turning to multi-dimensional distributions, simulation may be difficult. Gibbs sampling is a method specifically introduced to handle multi-dimensional distributions. Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ and let $\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)^T$, i.e. containing all components except X_i. Denote the univariate conditional density of $X_i|\mathbf{X}_{-i}$ by $f(x_i|\mathbf{x}_{-i})$, and suppose that it is easy to sample from these conditional densities. Gibbs sampling then uses these univariate and hopefully tractable conditional densities to construct the chain. A description of the algorithm can be found in Algorithm 2.2.

Algorithm 2.2 The Gibbs sampling algorithm. (Givens and Hoeting 2012)

1: Initialise $\mathbf{x}(0)$.
2: **Repeat** for all t:
3: Given $\mathbf{x}(t-1)$, generate, in sequence,

$$\begin{aligned} X_1(t)|\,(\cdot) &\sim f\left(x_1|\mathbf{x}_{-1}(t-1)\right) \\ X_2(t)|\,(\cdot) &\sim f\left(x_2|\mathbf{x}_{-2}(t-1)\right) \\ &\vdots \\ X_{d-1}(t)|\,(\cdot) &\sim f\left(x_{d-1}|\mathbf{x}_{-(d-1)}(t-1)\right) \\ X_d(t)|\,(\cdot) &\sim f\left(x_d|\mathbf{x}_{-d}(t-1)\right), \end{aligned} \tag{2.22}$$

where $\mathbf{x}_{-i}(t-1) = \big(x_1(t), \ldots, x_{i-1}(t), x_{i+1}(t-1), \ldots, x_d(t-1)\big)^T$, and where $(\cdot)$ represents conditioning on the most recent updates to all other components.
4: **End**

2.2 Stochastic differential equations

Here, we strive to give the reader a brief overview of the theory surrounding stochastic differential equations, as alluded to in the introduction. As with Langevin's century-old addition to Stoke's law, one may think of a diffusion process as consisting of one deterministic drift term, μ, guiding the process in a particular direction – regular fluid

dynamics, in Langevin's case – and one random term that disturbs the process' deterministic march, amplified by the factor σ – the aforementioned, erratic molecular movement. We now turn to finding an appropriate mathematical interpretation of the noise terms.

2.2.1 Mathematical preliminaries

Before one can define the – admittedly rather heavy – probabilistic machinery around which the theory of stochastic differential equations are built, one needs some mathematical foundations upon which to stand.

First, regard the problem of choosing random points uniformly from the interval $[0, 1)$, i.e. the possible outcomes are all points ω in the set $\Omega = [0, 1)$. Each outcome ought to be equiprobable, but since the interval in question is uncountable, the probability of each outcome is be zero. The idea that $\mathbb{P}(\omega) = 0$ for all $\omega \in [0, 1)$ is, however reasonable, useless. One might, of course, argue that the probability of obtaining a specific real number is irrelevant, and that one should rather be interested in the probability of obtaining a value in some interval A. One's intuition might then lead one to say that this probability is the sum of the probabilities of obtaining the individual points in the interval, i.e. that $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega)$. Once again, one has to admit to having created a rather pointless construction. Thus, a new mindset is needed. We therefore conclude that we must assign probabilities to subsets of the *sample space* Ω, rather that to individual outcomes. In this endeavour, the following definitions are useful, and constitute a brief introduction to probability. For a full treatment, the reader should consult e.g. Shiryayev and Boas 1984.

Definition 2.5. *Let Ω be a set of points ω. A system $\mathscr{A}$ of subsets of Ω is an* algebra *if*

i) $\Omega \in \mathscr{A}$

ii) $A, B \in \mathscr{A} \implies A \cup B \in \mathscr{A},\ A \cap B \in \mathscr{A}$

iii) $A \in \mathscr{A} \implies \bar{A} \in \mathscr{A}$,

where $\bar{A} = \Omega \setminus A$.

However, this class of subsets of Ω is too broad to yield a useful theory and needs to be restricted further, and we therefore introduce the concept of σ-algebras, which, in addition to being algebras, are completed to include countably infinite operations.

Definition 2.6. *A family $\mathscr{F}$ of subsets of Ω is a* σ-algebra *if it is an algebra and, furthermore, fulfils the following condition:*

ii)' if $A_n \in \mathscr{F},\ n = 1, 2, \ldots$, *then* $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$ *and* $\bigcap_{n=1}^{\infty} A_n \in \mathscr{F}$.

It should be noted that this is a stronger condition than condition *ii)* of Definition 2.5.

Definition 2.7. *The space Ω together with a σ-algebra $\mathscr{F}$ of its subsets constitutes a* measurable space*, denoted* $(\Omega, \mathscr{F})$.

Now that we have these set theoretical definitions (Shiryayev and Boas 1984), we are closer to a definition of probability. However, we still need some more results from measure theory.

Definition 2.8. *Let $\mathscr{G}$ be the collection of open subsets of a space Ω (e.g. $\Omega = \mathbb{R}^n$). Let $\mathscr{B}$ be the smallest σ-algebra that contains $\mathscr{G}$. $\mathscr{B}$ is then called the* Borel *algebra on Ω, and its sets are called* Borel sets. *(Øksendal 1998)*

Furthermore, we need the following definitions, the first one too broad, necessitating the second.

Definition 2.9. *A set function $\mu = \mu(A),\ A \in \mathscr{A}$, taking values in $[0, \infty]$, is called a* finitely additive measure *defined on $\mathscr{A}$ if*

$$\mu(A + B) = \mu(A) + \mu(B), \tag{2.23}$$

for every pair of disjoint sets A and B in $\mathscr{A}$. It is called finite if $\mu(\Omega) < \infty$.

Definition 2.10. *A finitely additive measure μ defined on an algebra $\mathscr{A}$ of subsets of Ω is σ-*additive *if, for all pairwise disjoint subsets $A_1, A_2, \ldots$ of A,*

$$\mu\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{2.24}$$

*It is said to be σ-*finite *if Ω can be represented on the form*

$$\Omega = \sum_{n=1}^{\infty} \Omega_n, \quad \Omega_n \in \mathscr{A}, \tag{2.25}$$

with $\mu(\Omega_n) < \infty,\ n = 1, 2, \ldots$

Definition 2.11. *A σ-additive measure $\mathbb{P}$ on $\mathscr{F}$ that satisfies $\mathbb{P}(\Omega) = 1$ is called a* probability measure.

(Shiryayev and Boas 1984)

Definition 2.12. *Let μ and ν be measures on the measurable space $(\Omega, \mathscr{F})$. The measure ν is said to be* absolutely continuous *with respect to μ on $\mathscr{F}$ if, for all $A \in \mathscr{F}$,*

$$\mu(A) = 0 \implies \nu(A) = 0, \tag{2.26}$$

which is written $\nu \ll \mu$. If it holds that both $\nu \ll \mu$ and $\mu \ll \nu$ then μ and ν are said to be equivalent*, which is written $\mu \sim \nu$. (Björk 2009)*

At last, we can set up a mathematical framework for probability with the following fundamental definition.

Definition 2.13. *An ordered triple $(\Omega, \mathscr{F}, \mathbb{P})$, where*

i) Ω is a set of points ω

ii) $\mathscr{F}$ is a σ-algebra of subsets of Ω

iii) $\mathbb{P}$ is a probability measure on $\mathscr{F}$,

is called a probabilistic model or probability space*. In this setting, Ω is denoted sample space, sets A in $\mathscr{F}$ are events, and $\mathbb{P}(A)$ is the probability of event A.*

(Shiryayev and Boas 1984)

Definition 2.14. *Given a probabilistic model $(\Omega, \mathscr{F}, \mathbb{P})$, a function $Y : \Omega \to \mathbb{R}^n$ is called* $\mathscr{F}$-measurable *if*

$$Y^{-1}(U) \coloneqq \{\omega \in \Omega;\ Y(\omega) \in U\} \in \mathscr{F}, \tag{2.27}$$

for all Borel sets $U \subset \mathbb{R}^n$. (Øksendal 1998)

We now return to Brownian motion, seeking a rigorous mathematical formulation. The random collisions of pollen grains with the molecules of a liquid observed by botanist Robert Brown in 1828 are conveniently modelled by a stochastic process $W(t)$. The interpretation hereof is the position of a pollen grain ω at time t. (Øksendal 1998)

Definition 2.15. *A stochastic process W is called a* Wiener process *(or standard Brownian motion) if it fulfils the following conditions.*

i) $W(0) = 0$*, almost surely.*

ii) The process W has independent increments, i.e. $W(t) - W(s)$ and $W(u) - W(v)$ are independent for $t > s \geq u > v$.

iii) The stochastic variable $W(t) - W(s),\ s < t$, has the distribution $N(0, t - s)$.

iv) W has continuous trajectories.

(Björk 2009)

Another important concept that often appears in the study of stochastic differential equations is that of martingales, which are defined as follows.

Definition 2.16. *A filtration on the measurable space $(\Omega, \mathscr{F})$ is a family $\mathscr{M} = \{\mathscr{M}_t\}_{t \geq 0}$ of σ-algebras $\mathscr{M}_t \subset \mathscr{F}$ such that $\{\mathscr{M}_t\}$ is increasing, i.e. that*

$$0 \leq s < t \implies \mathscr{M}_s \subset \mathscr{M}_t. \tag{2.28}$$

An n-dimensional stochastic process $\{M_t\}_{t \geq 0}$ on $(\Omega, \mathscr{F}, \mathbb{P})$ is called a martingale *with respect to a filtration $\{\mathscr{M}_t\}_{t \geq 0}$ if*

i) M_t is $\mathscr{M}_t$-measurable for all t

ii) $\mathbb{E}[|M_t|] < \infty$ for all t

iii) $\mathbb{E}[M_t | \mathscr{M}_s] = M_s$ for all $s \leq t$.

(Øksendal 1998)

2.2.2 Itō integrals

In order to write diffusion processes properly, one needs to rewrite (1.1) so that the driving random noise is a proper stochastic process, hereafter called $W(t)$. Regard, therefore, the equation at the discrete points $0 = t_0 < t_1 < \ldots < t_m = t$, with obvious notational simplifications,

$$X_{k-1} - X_k = \mu(t_k, X_k)\Delta t_k + \sigma(t_k, X_k)\Delta W_k, \tag{2.29}$$

where $\Delta t = t_{k+1} - t_k$ and $\Delta W_k = W(t_{k+1}) - W(t_k)$.

One assumes that the process $W(t)$ fulfils the following three assumptions:

i) $W(t_1)$ and $W(t_2) - W(t_1)$ are independent for $t_1 \leq t_2$.

ii) $\{\Delta W(t)\}_{t \geq 0}$ is stationary.

iii) $\mathbb{E}[W_t] = 0 \quad \forall\, t$.

By these assumptions, $W(t)$ should have stationary independent increments with zero mean, and the only stochastic process with continuous paths for which this is true is the Brownian motion. Hence, one uses the Brownian motion as driving noise for diffusion processes. Starting from some initial value X_0, one may sum the individual terms, obtaining

$$X_k = X_0 + \sum_{j=0}^{k-1} \mu(t_j, X_j)\Delta t_j + \sum_{j=0}^{k-1} \sigma(t_j, X_j)\Delta W_j. \tag{2.30}$$

Taking the limit of (2.30) as $\Delta t_j \to 0$, the existence of which is possible to prove, yields

$$X_k = X_0 + \int_0^t \mu(s, X_s)ds + \int_0^t \sigma(s, X_s)dW. \tag{2.31}$$

By convention, one says that (1.2) really means that $X(t)$ is a stochastic process that satisfies this equation, (2.31).

This argument, albeit nice at a first glance, raises the question: What does the second integral really mean? Thus, it is now necessary to define the integral

$$\int_a^b g(s)dW(s), \tag{2.32}$$

for some suitable class of functions g. The 'suitable' class is defined as follows.

Definition 2.17. *Let $\mathscr{B}$ denote the Borel σ-algebra on $[0, \infty)$. Then let $\mathcal{L}^2[a, b]$ be the class of functions $g(t, \omega) : [0, \infty) \times \Omega \to \mathbb{R}$ such that*

i) $(t, \omega) \to g(t, \omega)$ is $\mathscr{B} \times \mathscr{F}$-measurable

ii) $g(t, \omega)$ is $\mathscr{F}_t$-adapted

iii) $\mathbb{E}\left[\int_a^b g^2(t, \omega)dt\right] < \infty$.

(Def. 3.1.4, Øksendal 1998)

Furthermore, one says that a process g belongs to the aforementioned class $\mathcal{L}^2$ if $g \in \mathcal{L}^2[0, t] \quad \forall\, t > 0$.

It is sensible to begin by regarding *simple* functions $g \in \mathcal{L}^2[a, b]$, i.e. functions that have deterministic points in time, $a = t_0 < t_1 < \ldots < t_n = b$, between which the function g is constant. The obvious definition is then

$$\int_a^b g(s)dW(s) = \sum_{k=0}^{n-1} g(t_k)[W_{k+1} - W_k], \tag{2.33}$$

where one looks at *forward* increments of the Brownian motion.

For general, that is, not simple, processes $g \in \mathcal{L}^2[a, b]$, one may approximate the process to a sequence of simple processes g_n, treated above, such that

$$\int_a^b \mathbb{E}\left[\{g_n(s) - g(s)\}^2\right] ds \to 0 \quad \text{as } n \to \infty. \tag{2.34}$$

The integral $\int_a^b g_n(s)dW(s)$ is then a well defined stochastic variable Z_n for each n, and it is possible to prove, although far outside the scope of this treatment, that there exists a stochastic variable Z such that $Z_n \to Z$ in $\mathscr{L}^2$ as $n \to \infty$. Thus, one defines the stochastic integral as follows.

Definition 2.18. *Let $g \in \mathcal{L}^2[a,b]$. The* Itō integral *is then defined as*

$$\int_a^b g(s)dW(s) = \lim_{n\to\infty} \int_a^b g_n(s)dW(s). \tag{2.35}$$

(Björk 2009)

It should be noted that this is not the only way of defining a stochastic integral. Another definition is the Stratonovich integral, where, instead of using forward increments, one uses mid-points. (Øksendal 1998)

The Itō integral has the following practical properties.

Corollary 2.19. Itō isometry. *Let $g \in \mathcal{L}^2[a,b]$, then*

$$\mathbb{E}\left[\left(\int_a^b g(t,\omega)dW(t)\right)^2\right] = \mathbb{E}\left[\int_a^b g^2(t,\omega)dt\right]. \tag{2.36}$$

Theorem 2.20. *Let $f, g \in \mathcal{L}^2[a,b]$, $0 \le a < u < b$, and let c be a constant. Then*

i) $\int_a^b g(t,\omega)dW(t) = \int_a^u g(t,\omega)dW(t) + \int_u^b g(t,\omega)dW(t)$ *for almost every ω*

ii) $\int_a^b \big(cf(t,\omega)+g(t,\omega)\big)dW(t) = c\int_a^b f(t,\omega)dW(t) + \int_a^b g(t,\omega)dW(t)$ *for almost every ω*

iii) $\mathbb{E}\left[\int_a^b g(t,\omega)dW(t)\right] = 0$

iv) $\int_a^b g(t,\omega)dW(t)$ *is $\mathscr{F}_b^W$ measurable.*

(Øksendal 1998) Furthermore, it follows that

Corollary 2.21. *For any process $g \in \mathcal{L}[0,t]$, the process*

$$X(t) = \int_0^t g(s)dW(s), \tag{2.37}$$

i.e. any stochastic integral where $g \in \mathcal{L}^2$, is an $\mathscr{F}_t^W$-martingale. (Corollary 4.8, Björk 2009)

Proof. For fixed s and t such that $s < t$, and by *i)* in theorem 2.20, we have that

$$\begin{aligned}\mathbb{E}\left[X(t)|\mathscr{F}_s^W\right] &= \mathbb{E}\left[\int_0^t g(\tau)dW(\tau)|\mathscr{F}_s^W\right] \\ &= \mathbb{E}\left[\int_0^s g(\tau)dW(\tau)|\mathscr{F}_s^W\right] + \mathbb{E}\left[\int_s^t g(\tau)dW(\tau)|\mathscr{F}_s^W\right],\end{aligned} \tag{2.38}$$

where the first term, by *iv)* in theorem 2.20, is $\mathscr{F}_s^W$-measurable, and the second term, by *iii)* in theorem 2.20, equals zero. Thus,

$$\mathbb{E}\left[X(t)|\mathscr{F}_s^W\right] = \int_0^s g(\tau)dW(\tau) + 0 = X(s), \tag{2.39}$$

or, in other words, $X(t)$ is a martingale. (Björk 2009) □

In fact, there is an even stronger result that is very useful.

Lemma 2.22. *Assuming sufficient integrability and continuous paths, a stochastic process X is a martingale if and only if its stochastic differential is of the form*

$$dX(t) = g(t)dW(t), \tag{2.40}$$

i.e. it has no drift term. (Lemma 4.9, Björk 2009)

It is possible to extend this definition of the Itō integral to a larger class of functions than $\mathcal{L}^2$. By relaxing the measurability condition of Definition 2.17 to:

i)' There exists and increasing family of σ-algebras $\mathcal{H}_t;\ t \geq 0$ such that

   a. W_t is a martingale with respect to $\mathcal{H}_t$

   b. $g(t)$ i $\mathcal{H}_t$-adapted.

Thus, we can allow g to depend on more than just $\mathcal{F}_t$, which is a subset of the new family, $\mathcal{F}_t \subset \mathcal{H}_t$, as long as the Wiener process $W(t)$ remains a martingale with respect to the history of the process g. The above condition means that $\mathbb{E}[B_t - B_s|\mathcal{H}_s] = 0$ for all $s < t$, which is actually sufficient to define the Itō integral as in Definition 2.18.

This relaxation is necessary to extend the Itō integral to multiple dimensions. Let $W = (W_1, \ldots, W_n)$ be an n-dimensional Brownian motion, and let $\mathcal{F}_t^{(n)}$ the σ-algebra generated by all the coordinates of W. Since $W_k(t) - W_k(s)$ is independent of $\mathcal{F}_s^{(n)}$ when $s < t$, W_k is a martingale with regards to $\mathcal{F}_s^{(n)}$. Thereby, the integral of some $\mathcal{F}_s^{(n)}$-adapted function g with regards to a coordinate of W may be defined.

Definition 2.23. *Let $W = (W_1, \ldots, W_n)$ be an n-dimensional Brownian motion, and let $\mathcal{L}^{2,\mathcal{H}}_{m\times n}[a,b]$ denote the set of $m \times n$ matrices wherein each element $g_{ij}(t,\omega)$ satisfies the condition i)' above and the last two conditions of Definition 2.17 with regards to some filtration $\mathcal{H} = \{\mathcal{H}_t\}_{t\geq 0}$. Then define the multi-dimensional Itō integral as the $m \times 1$ vector*

$$\int_a^b g(t,\omega)dW = \int_a^b \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \ddots & \vdots \\ g_{m1} & \cdots & g_{mn} \end{pmatrix} \begin{pmatrix} dW_1 \\ \vdots \\ dW_n \end{pmatrix}, \tag{2.41}$$

whose ith component is the sum of unidimensional Itō integrals, extended as above:

$$\sum_{j=1}^{n} \int_a^b g_{ij}(s,\omega)dW_j(s,\omega). \tag{2.42}$$

(Øksendal 1998)

2.2.3 The Itō formula

After having set up the framework of the previous section, it is with some regret that one notices that the basic definition of the Itō integral is often impractical when evaluating given integrals. Fortunately, it is possible to establish a formula for the Itō integral that serves much the same purpose as the chain rule does in regular calculus. It is called the Itō formula, after its creator Kiyosi Itō . We start by looking at an example.

Example 2.24. Consider the stochastic integral $\int_0^t W(s)dW(s)$, where, for simplicity, $W(0) = 0$. Following the definition, we set $g_n(s) = \sum W(t_j)\mathbb{1}_{\{t_j,t_{j+1}\}}(s)$, where $\mathbb{1}$ is the indicator function. We then have

$$\begin{aligned}\mathbb{E}\left[\int_0^t (g_n(s) - W(s))^2 ds\right] &= \mathbb{E}\left[\sum_j \int_{t_j}^{t_{j+1}} (W(t_j) - W(s))^2 ds\right] \\ \sum_j \int_{t_j}^{t_{j+1}} (s - t_j) ds &= \sum_j \frac{1}{2}(t_{j+1} - t_j)^2 \to 0, \quad \text{as } \Delta t_j \to 0.\end{aligned} \tag{2.43}$$

Thus, by Definition 2.18,

$$\int_0^t W(s)dW(s) = \lim_{\Delta t_j \to 0} \int_0^t g_n dW(s) = \lim_{\Delta t_j \to 0} \sum_j W_j \Delta W(t_j). \tag{2.44}$$

Furthermore,

$$\begin{aligned}\Delta(W^2(t_j)) &= W^2(t_{j+1}) - W^2(t_j) \\ &= (W(t_{j+1}) - W(t_j))^2 + 2W(t_j)(W(t_{j+1}) - W(t_j)) \\ &= (\Delta W(t_j))^2 + 2W(t_j)\Delta W(t_j),\end{aligned} \tag{2.45}$$

whereby, since $W(0) = 0$,

$$W(t)^2 = \sum_j \Delta(W^2(t_j)) = \sum_j (\Delta W(t_j))^2 + 2\sum_j W(t_j)\Delta W(t_j), \tag{2.46}$$

or, equivalently,

$$\sum_j W(t_j)\Delta W(t_j) = \frac{1}{2}W^2(t) - \frac{1}{2}\sum_j (\Delta W(t_j))^2. \tag{2.47}$$

Now, turning to the last term, we have, since $t_j = j\frac{t}{n}$, where n is the number of sub-intervals, that

$$\begin{aligned}\mathbb{E}\left[\sum_j (\Delta W(t_j))^2\right] &= \sum_j \mathbb{E}\left[(W(t_{j+1}) - W(t_j))^2\right] \\ &= \sum_j \left[(j+1)\frac{t}{n} - j\frac{t}{n}\right] = t,\end{aligned} \tag{2.48}$$

and, as W has independent increments, that

$$\begin{aligned}\mathbb{V}\left[\sum_j (\Delta W(t_j))^2\right] &= \sum_j \mathbb{V}\left[(W(t_{j+1}) - W(t_j))^2\right] \\ &= \sum_j 2\left[\frac{t^2}{n^2}\right] = 2\frac{t^2}{n} \to 0 \quad \text{as } \Delta t_j \to 0.\end{aligned} \tag{2.49}$$

Thereby, $\sum_j (\Delta W(t_j))^2$ tends to the deterministic limit t (in L^2) as $n \to \infty$, or, equivalently, as $\Delta t_j \to 0$. Hence, we find that

$$\begin{aligned}\int_0^t W(s)dW(s) &= \frac{1}{2}W^2(t) - \lim_{\Delta t_j \to 0}\left(\frac{1}{2}\sum_j (\Delta W(t_j))^2\right) \\ &= \frac{1}{2}W^2(t) - \frac{1}{2}t.\end{aligned} \tag{2.50}$$

This example shows not only that the Itō integral does not behave like 'regular' integrals, but also that the definition is cumbersome to work with. (Øksendal 1998)

The image of the Itō integral by a map $g(x)$ does not yield an Itō integral alone, but also a 'ds-integral'. It is therefore natural to define the (one dimensional) *Itō process* as follows. For notational convenience, a process at time t is henceforth denoted as both $X(t)$ and X_t.

Definition 2.25. *Let W_t be a one-dimensional Wiener process on $(\Omega, \mathscr{F}, \mathbb{P})$. A one-dimensional* Itō process *is then a stochastic process X_t on $(\Omega, \mathscr{F}, \mathbb{P})$ of the form*

$$X_t = X_0 + \int_0^t \mu(s,\omega)ds + \int_0^t \sigma(s,\omega)dW_s, \tag{2.51}$$

or, written on a more compact form,

$$\begin{aligned} dX_t &= \mu(t, X_t)dt + \sigma(t, X_t)dW_t, \\ X_0 &= x_0, \end{aligned} \tag{2.52}$$

where μ and σ are adapted processes. This last form is called the stochastic differential *of X.*

Now, we state the most important result of Itō calculus, the famous and for practical purposes very useful Itō formula.

Lemma 2.26. *Let $dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t$ be an Itō process, and let $g(t,x) : [0,\infty) \times \mathbb{R} \to \mathbb{R}$ be a $C^{1,2}$-function, i.e. once differentiable with regards to the first argument and twice differentiable with regards to the second argument. Then*

$$Y_t = g(t, X_t) \tag{2.53}$$

is an Itō process with the stochastic differential

$$dY_t = \frac{\partial}{\partial t}g(t, X_t)dt + \frac{\partial}{\partial x}g(t, X_t)dX_t + \frac{1}{2}\frac{\partial^2}{\partial x^2}g(t, X_t)(dX_t)^2. \tag{2.54}$$

The term $(dX_t)^2 = dX_t \cdot dX_t$ is calculated according to the rules

$$dt \cdot dt = dt \cdot dW_t = dW_t \cdot dt = 0, \quad dW_t \cdot dW_t = dt. \tag{2.55}$$

In higher dimensions, the result is much the same. To examine the d-dimensional vector X, one regards instead the Wiener process $W(t) = [W_1(t), \ldots, W_m(t)]^T$, with m dimensions, the d-dimensional drift vector $\mu = [\mu_1, \ldots, \mu_d]^T$, and the $d \times m$-dimensional diffusion matrix

$$\sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dm} \end{bmatrix}, \tag{2.56}$$

with similar conditions on μ and σ as in the one-dimensional case. The question of applying a smooth function on the d-dimensional Itō process X gives rise to the following multi-dimensional version of Itō's formula.

Theorem 2.27. *Let $dX(t) = \mu(t, X_t)dt + \sigma(t, X_t)dW(t)$ be a multi-dimensional Itō process as above, and let $g(t, \mathbf{x}) : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^p$ be a $C^{1,2}$-map. Then*

$$Y(t) = g(t, X_t) \tag{2.57}$$

is an Itō process, the kth component of which is given by the stochastic differential

$$dY_k = \frac{\partial}{\partial t}g_k(t, X)dt + \sum_i \frac{\partial}{\partial x_i}g_k(t, X_i)dX + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}g_k(t, X)dX_i dX_j, \tag{2.58}$$

where, for notational brevity, the dependence on time of the stochastic processes has been omitted. For calculations, the following rules apply.

$$dt \cdot dt = dt \cdot dW_i(t) = dW_i(t) \cdot dt = 0, \quad dW_i(t) \cdot dW_j(t) = \delta_{ij}dt, \tag{2.59}$$

where δ_{ij} is the Kronecker delta.

(Øksendal 1998)

There are many connections between Itō processes and partial differential equations, a connection that is often extensively used in applications. An Itō diffusion X_t is associated with a partial differential operator $\mathscr{A}$, and one says that $\mathscr{A}$ is the generator of X_t.

Definition 2.28. *Let $\{X_t\}$ be a time homogeneous Itō process in $\mathbb{R}^n$. The* infinitesimal generator *$\mathscr{A}$ of X_t is then defined by*

$$\mathscr{A}f(x) = \lim_{t \downarrow 0} \frac{\mathbb{E}^x[f(X_t)] - f(x)}{t}, \quad x \in \mathbb{R}^n. \tag{2.60}$$

The set of functions $f : \mathbb{R}^n \to \mathbb{R}$, such that this limit exists at x is denoted $\mathscr{D}_A(x)$, and the set of functions for which the limit exists for all $x \in \mathbb{R}^n$ is denoted $\mathscr{D}_A$.

Now, this definition may seem rather abstruse, and some tangible formula is needed.

Theorem 2.29. *Let X_t be the Itō diffusion*

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t. \tag{2.61}$$

If the function f has compact support and is twice differentiable, i.e. $f \in \mathscr{C}_0^2(\mathbb{R}^n)$, we have that

$$\mathscr{A}f(x) = \sum_i \mu_i(x)\frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j}(\sigma\sigma^T)_{i,j}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}. \tag{2.62}$$

Furthermore, we have the following important result.

Theorem 2.30. *Let $f \in \mathscr{C}_0^2(\mathbb{R}^n)$, and define*

$$u(t, x) = \mathbb{E}^x[f(x)]. \tag{2.63}$$

Then $u(t,\cdot) \in \mathscr{D}_A$ for each t and

$$\begin{aligned} \frac{\partial u}{\partial t} &= \mathscr{A}u, \quad t > 0,\ x \in \mathbb{R}^n, \\ u(0,x) &= f(x). \end{aligned} \tag{2.64}$$

Furthermore, if $w \in \mathscr{C}^{1,2}(\mathbb{R} \times \mathbb{R}^n)$ is a bounded function that satisfies the aforementioned equations, then $w(t,x) = u(t,x)$.

If one takes the function f to be the indicator function $\mathbb{1}_{X_T \in B}$, one obtains the probability $U(s,y) = \mathbb{P}[X_T \in B | X_s = y]$. One can then deduce that the following relation holds for the transition probabilities of an SDE.

$$\begin{gathered} \left(\frac{\partial p}{\partial t} + \mathscr{A}u\right)(s,y;t,x) = 0, \quad t > 0,\ x \in \mathbb{R}^n, \\ p(s,y;t,x) \to \delta_x, \quad \text{as } s \to t. \end{gathered} \tag{2.65}$$

This is known as the Kolmogorov backward equation as the operator is working on the 'backward' variables (s,y), rather than the 'forward' variables (t,x). One can also derive a corresponding forward equation, here given only for transition densities.

Theorem 2.31. *Let $f \in \mathscr{C}_0^2(\mathbb{R}^n)$, and define*

$$u(t,x) = \mathbb{E}^x[f(x)]. \tag{2.66}$$

Then $u(t,\cdot) \in \mathscr{D}_A$ for each t and

$$\begin{aligned} \frac{\partial p}{\partial t}(s,y;t,x) &= \mathscr{A}^* p(s,y;t,x), \quad t > 0,\ x \in \mathbb{R}^n, \\ p(s,y;t,x) &\to \delta_y, \quad \text{as } t \downarrow s. \end{aligned} \tag{2.67}$$

$\mathscr{A}^*$ is the adjoint operator of $\mathscr{A}$ and is given by

$$\mathscr{A}^* f(t,x) = -\sum_i \frac{\partial}{\partial x_i}\mu_i(t,x) f(t,x) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}(\sigma\sigma^T)_{i,j}(t,x) f(t,x). \tag{2.68}$$

The Kolmogorov forward equation is sometimes called the *Focker-Planck equation*. (Øksendal 1998) It has furthermore been suggested that the Fokker-Planck equation be used as a basis for parameter estimation, see e.g. Lo 1988 and Erik Lindström 2007

2.2.4 Solutions to stochastic differential equations

As the previous sections have shown, there exists a rich theory on Itō processes and stochastic differential equations (SDEs), and it is therefore natural to pose the question: Do solutions to these stochastic differential equations exist, and are they unique?

If the diffusion $\sigma(t, X_t)$, in (1.2), is identically equal to zero, the stochastic differential equation reduces to an ordinary differential equation. It is well known from regular calculus that a unique solution exists if the drift μ is Lipschitz continuous in the space variable X_t (the Picard-Lindelöf theorem). It is therefore natural to consider similar, Lipschitz-type conditions on stochastic differential equations when examining the existence and uniqueness of their solutions, something that was first done by Kiyosi Itō. We have the following theorem for the existence and uniqueness in SDEs, presented here without proofs. For a more complete exposition, see e.g. Karatzas and Shreve 1998.

Theorem 2.32. *Strong uniqueness holds for* (1.2) *if the coefficients* $\mu(t,x)$ *and* $\sigma(t,x)$ *are locally Lipschitz continuous in the space variable, i.e. that there, for every integer* $n \geq 1$*, exists a constant* $K_n > 0$ *such that*

$$|\mu(t,x) - \mu(t,y)| + |\sigma(t,x) - \sigma(t,y)| \leq K_n|x-y|, \tag{2.69}$$

for every $t \geq 0$*,* $|x| \leq n$*, and* $|y| \leq n$*.*

This is, however, not enough to ensure the existence of a unique solution for all t. In order for that to be the case, we need the additional assumption of linear growth of the coefficients, i.e.

$$|\mu(t,x)| + |\sigma(t,x)| \leq C(1+|x|) \quad \forall x \in \mathbb{R}, \tag{2.70}$$

where C is a positive, real constant.

In one dimension, these conditions may be relaxed considerably. This is fortunate as there are stochastic differential equations that have a solution, but that do not fulfil these conditions. A prominent example is the Cox-Ingersoll-Ross model, which is commonly used in finance, but whose diffusion is not Lipschitz continuous.

Theorem 2.33. *Suppose that* $\mu(t,x)$ *and* $\sigma(t,x)$ *satisfy the conditions*

$$|\mu(t,x) - \mu(t,y)| \leq \kappa(|x-y|), \tag{2.71}$$

$$|\sigma(t,x) - \sigma(t,y)| \leq h(|x-y|), \tag{2.72}$$

for every $0 \leq t < \infty$ *and* $x \in \mathbb{R},\ y \in \mathbb{R}$*, where* $h : [0,\infty) \to [0,\infty)$ *is a strictly increasing function with* $h(0) = 0$ *and*

$$\int_{(0,\varepsilon)} h^{-2}(u)du = \infty, \quad \forall \varepsilon > 0, \tag{2.73}$$

and $\kappa : [0,\infty) \to [0,\infty)$ *is a strictly increasing and concave function with* $\kappa(0) = 0$ *and*

$$\int_{(0,\varepsilon)} \kappa^{-1}(u)du = \infty, \quad \forall \varepsilon > 0, \tag{2.74}$$

Then there exists a unique solution to (1.2)*.*

It is also worth noting that the solution is a Markov process. (Karatzas and Shreve 1998)

Moving on, it is instructive to regard some examples of stochastic differential equations and their solutions, starting with the geometric Brownian motion, which is of great importance in e.g. finance, as it forms the basis for the famous Black-Scholes model.

Example 2.34. *Geometric Brownian motion.* Consider the simple stochastic differential equation

$$\begin{aligned} dX_t &= \alpha X_t dt + \sigma X_t dW_t, \\ X_0 &= x_0. \end{aligned} \tag{2.75}$$

If one, for a second, disregards the stochastic term, and looks at the corresponding ordinary differential equation (ODE), it is known from early courses in calculus that the solution hereto is an exponential. With inspiration from this, one may examine

the auxiliary process $Z_t = \log X_t$, under the assumption that X_t is a strictly positive solution to (2.75). Itō's formula then gives the dynamics of the auxiliary process as

$$\begin{aligned} dZ_t &= \frac{1}{X_t} dX_t + \frac{1}{2}(dX_t)^2 \\ &= \frac{1}{X_t}[\alpha X_t dt + \sigma X_t dW_t] + \frac{1}{2}\left(-\frac{1}{X_t^2}\right)\sigma^2 X_t^2 dt \\ &= \left(\alpha + -\frac{1}{2}\sigma^2\right) dt + \sigma dW_t, \end{aligned} \tag{2.76}$$

with $Z_0 = \log x_0$. Since this does not contain any Z_t, the equation is easy to solve by direct integration, and one obtains $Z_t = \log x_0 + \left(\alpha - \frac{1}{2}\sigma^2\right) t + \sigma W_t$. This, in turn, means that X_t is obtained as

$$X_t = x_0 \exp\left\{\left(\alpha - \frac{1}{2}\sigma^2\right) t + \sigma W_t\right\}. \tag{2.77}$$

It should be noted that the assumption of strict positivity is somewhat problematic. Therefore, technically, one defines the solution to (2.75) as the recently obtained (2.77), dodging the aforementioned problem. It is easy to show that this defined solution actually satisfies the stochastic differential equation in question. (Björk 2009)

Example 2.35. *The linear SDE.* Next, consider the following stochastic differential equation:

$$\begin{aligned} dX_t &= (\alpha X_t + u_t)dt + \sigma_t dW_t, \\ X_0 &= x_0, \end{aligned} \tag{2.78}$$

where u_t is a deterministic function of time. If one, once again, starts by looking at the corresponding ordinary differential equation $\dot{x}_t = \alpha x_t + u_t$, one finds that the solution to the ODE is:

$$x_t = e^{\alpha t} x_0 + \int_0^t e^{\alpha(t-s)} u_s ds. \tag{2.79}$$

As an analogue, it is tempting to conjecture that the solution to the SDE is similar to the deterministic solution, and that the stochastic term enters in much the same way as the deterministic function does in the ODE solution. This is, of course, a purely heuristic argument, but it turns out that a solution defined in this manner,

$$X_t = e^{\alpha t} x_0 + \int_0^t e^{\alpha(t-s)} u_s ds + \int_0^t e^{\alpha(t-s)} \sigma dW_s, \tag{2.80}$$

does indeed satisfy the linear SDE. (Björk 2009) If one takes u_t and σ_t to be constants, and take the parameter α to be negative, one obtains the well-known Ornstein-Uhlenbeck process.

Example 2.36. *Cox-Ingersoll-Ross model (CIR).* Introduced by Cox, Ingersoll Jr and Ross 1985, this is a model that one often encounters in finance. It is given by the following SDE:

$$dX_t = \kappa(\theta - X_t)dt + \sigma\sqrt{X_t} dW_t, \quad X_0 = x_0. \tag{2.81}$$

The diffusion of this common model, to which a solution exists, is not Lipschitz continuous around zero, thus not fulfilling the conditions for existence and uniqueness stipulated in Theorem 2.32. However, this can be remedied by the following lemma.

Lemma 2.37. *The function $f(x) = \sigma\sqrt{x}$, defined for positive σ on the non-negative real line, is Hölder continuous of order $\frac{1}{2}$, i.e.*

$$|\sigma\sqrt{x} - \sigma\sqrt{y}| \leq C \parallel x - y \parallel^{\frac{1}{2}}, \tag{2.82}$$

for all non-negative x and y.

Proof. Since the function $f(x) = \sigma\sqrt{x}$ is non-negative, it is clear that

$$|\sigma\sqrt{x} - \sigma\sqrt{y}| \leq |\sigma\sqrt{x} + \sigma\sqrt{y}|. \tag{2.83}$$

Hence, we have that

$$|\sigma\sqrt{x} - \sigma\sqrt{y}|^2 \leq |\sigma\sqrt{x} - \sigma\sqrt{y}| \cdot |\sigma\sqrt{x} + \sigma\sqrt{y}| = |\sigma^2 x - \sigma^2 y|, \tag{2.84}$$

and the desired result is thus obtained as

$$|\sigma\sqrt{x} - \sigma\sqrt{y}| \leq \sigma|x - y|^{\frac{1}{2}}, \tag{2.85}$$

i.e. $f(x) = \sigma\sqrt{x}$ is Hölder continuous of order $\frac{1}{2}$. □

Thus, by Lemma 2.37, the conditions of Theorem 2.33 are fulfilled, and a unique solutions exists.

For the Cox-Ingersoll-Ross model, one can even obtain a closed form expression for the transition density as

$$p(X_s, s; X_t, t) = c\, e^{-u-v} \left(\frac{v}{u}\right)^{q/2} I_q(2\sqrt{uv}), \tag{2.86}$$

where

$$\begin{aligned} c &\equiv \frac{2\kappa}{\sigma^2(1 - e^{-\kappa(s-t)})}, \\ u &\equiv cX_t e^{-\kappa(s-t)}, \\ v &\equiv cX_t, \\ q &\equiv \frac{2\kappa\theta}{\sigma^2} - 1, \end{aligned} \tag{2.87}$$

and $I_q(\cdot)$ is the modified Bessel function of the first kind of order q. The transition density is obtained by solving the Fokker-Planck equation using Laplace transforms. This is carried out, albeit on a slightly different form, in Feller 1951. The transitions of the CIR model are thus distributed according to a non-central χ^2-distribution.

Example 2.38. *Lorenz model*. Consider the stochastic Lorenz model, which is given in three dimensions as:

$$\begin{bmatrix} dX_t^{(1)} \\ dX_t^{(2)} \\ dX_t^{(3)} \end{bmatrix} = \begin{bmatrix} s(X_t^{(2)} - X_t^{(1)}) \\ rX_t^{(1)} - X_t^{(2)} - X_t^{(1)} X_t^{(3)} \\ X_t^{(1)} X_t^{(2)} - bX_t^{(3)} \end{bmatrix} dt + \sigma \begin{bmatrix} dW_t^{(1)} \\ dW_t^{(2)} \\ dW_t^{(3)} \end{bmatrix}, \tag{2.88}$$

where s, r, b, and σ are parameters. Common choices for parameters are $[s,\ r,\ b,\ \sigma] = [10,\ 28,\ 8/3,\ 2]$, see e.g. Bengtsson, Snyder and Nychka 2003 or Erik Lindström 2012.

In its deterministic form, the Lorenz model was originally proposed to represent 'forced dissipative hydrodynamic flow'. (Lorenz 1963) For $r \geq 1$, the model creates a solution that orbits two attractors, and, for $r < 1$, a single attractor.

The drift does not conform to the linear growth condition and the regular theorems for existence and uniqueness do not apply. However, it is shown in Keller 1996 that the stochastic Lorenz system does indeed possess a unique solution.

2.2.5 Diffusion bridges

Consider the process X conditioned not only on the initial value u but also some ending point v. The unconditioned process has the SDE:

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t. \tag{2.89}$$

It is not straight-forward to find out how the dynamics of this new *Bridge process* will look. This section is devoted to that endeavour.

Denote the conditional density of X at time t, conditioned on starting from u at $t = 0$ and on hitting the point v at time T, by $p(x, t|u, 0, v, T)$, and, using the Markov property, rewrite it as

$$p(x, t|u, 0, v, T) = \frac{p(v, T|x, t)p(x, t|u, 0)}{p(v, T|u, 0)} = \frac{p \cdot p(x, t|u, 0)}{\tilde{p}}. \tag{2.90}$$

Now, the right hand side does not contain densities conditioned on a future event. For nice functions $f(x)$, one has that

$$\begin{aligned} \mathbb{E}[f(X_t)|X_0 = u, X_T = v] &= \int f(x)p(x, t|u, 0, v, T)dx \\ &= \int f(x)\frac{p(v, T|x, t)p(x, t|u, 0)}{p(v, T|u, 0)}dx \\ &= \frac{\mathbb{E}[f(x)p(v, T|x, t)|X_0 = u]}{p(v, T|u, 0)} \end{aligned} \tag{2.91}$$

Now, denote $g(t, x) = f(x)p(v, T|x, t)$, use Itō's formula as usual, and take the expectation, whereby one obtains

$$\mathbb{E}[g(t, X_t)] = \mathbb{E}\left[\int_0^t \left(\mu\frac{\partial}{\partial x}g(s, X_s) + \frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2}g(s, X_s) + \frac{\partial}{\partial t}g(s, X_s)\right)ds\right]. \tag{2.92}$$

Expanding and re-arranging gives, with arguments suppressed for notational brevity and subscripts denoting partial derivatives,

$$\mathbb{E}[g(t, X_t)] = \mathbb{E}\left[\int_0^t \left(\mu p f_x + \frac{1}{2}\sigma^2 p f_{xx} + \sigma^2 p_x f_x + f(\mu p_x + \frac{1}{2}\sigma^2 p_{xx} + p_s)\right)ds\right], \tag{2.93}$$

where the last parenthesis of the integrand equals zero by the Kolmogorov backward equation, Theorem 2.30. Thus, with $\tilde{p} = p(v, T|u, 0)$, one obtains

$$\mathbb{E}[f(X_t)|X_0 = u, X_T = v] = \frac{1}{\tilde{p}}\mathbb{E}\left[\int_0^t \left((\mu p + \sigma^2 p_x)f_x + \frac{1}{2}\sigma^2 p f_{xx}\right)ds|X_0 = u\right]. \tag{2.94}$$

Differentiation of both sides with respect to t, and conversion back to the full condi-

tional density, dividing by $p = p(v, T|x, t)$ above and below as needed, yields

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}[f(X_t)|X_0 = u, X_T = v] &= \mathbb{E}\left[\frac{p\left((\mu + \frac{\sigma^2}{p}p_x)f_x + \frac{1}{2}\sigma^2 f_{xx}\right)}{\tilde{p}}|X_0 = u\right] \\
&= \int \left[(\mu + \frac{\sigma^2}{p}p_x)f_x + \frac{1}{2}\sigma^2 f_{xx}\right]\frac{p}{\tilde{p}}p(x, t; u, 0)\, dx \\
&= \int \left[(\mu + \frac{\sigma^2}{p}p_x)f_x + \frac{1}{2}\sigma^2 f_{xx}\right]p(x, t; u, 0, v, T)\, dx \\
&= \mathbb{E}\left[(\mu + \frac{\sigma^2}{p}p_x)f_x + \frac{1}{2}\sigma^2 f_{xx}|X_0 = u, X_T = v\right] \\
&= \mathbb{E}\left[(\mu + \sigma^2\frac{\partial}{\partial x}\log p)\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 f}{\partial x^2}|X_0 = u, X_T = v\right],
\end{aligned} \tag{2.95}$$

where the last equality comes from applying the chain rule on the logarithmic function. Hence, one can deduce the Focker-Planck equation by using integration by parts and obtain the SDE of the bridge process. The result is stated as a theorem.

Theorem 2.39. *Consider the stochastic differential equation* (2.89)*. The SDE of a the corresponding bridge process, i.e. the process conditioned to hit a final value $X_T = v$, is given by*

$$dX_t = \left(\mu + \sigma^2\frac{\partial}{\partial x}\log p(v, T|X_t, t)\right)dt + \sigma dW_t. \tag{2.96}$$

(Lyons 2013)

The conditioned process, somewhat counter-intuitively, retains the Markov property, and the diffusion coefficient remains unchanged, whereas the drift acquires an extra term that forces the process to hit the end-point. The stochastic differential equation for the bridge process is insightful but typically intractable for all but the most simple models. Application of regular approximation techniques requires knowledge of the transition density, which one generally does not have. One therefore has to resort to alternative methods for simulating diffusion bridges, often Monte Carlo methods. (Papaspiliopoulos and G. Roberts 2012)

Example 2.40. *Brownian bridge.* A Brownian bridge is, quite simply, a Brownian motion from some point $\tilde{x}_0$, conditioned to end at some point $\tilde{x}_T$. Starting from $dX_t = \mu dt + \sigma dW_t$, it is an easy application of Theorem 2.39 to see that the SDE of the bridge is given by

$$d\tilde{X}_t = \frac{\tilde{x}_T - \tilde{X}_t}{T - t}dt + d\tilde{W}_t. \tag{2.97}$$

Equation (2.97) being essentially a linear SDE, we glance at Example 2.35 and proceed by the same arguments as given therein, setting $\alpha = -1/(T - t)$, $u_t = \tilde{x}_T/(T - t)$, and $\sigma_t = 1$. Hence, we obtain the explicit solution as

$$\tilde{x}_t = \tilde{x}_0\left(1 - \frac{t}{T}\right) + \tilde{x}_T\frac{t}{T} + (T - t)\int_0^t \frac{1}{T - s}d\tilde{W}_s. \tag{2.98}$$

Although a process is a Brownian bridge only if its diffusion is constant, it is possible to construct an analogous sampler for processes with non-constant diffusions in a manner similar to the Euler-Maruyama scheme. We set

$$\tilde{x}_{m+1} = \tilde{x}_m + \tilde{\mu}(\tilde{x}_m, \tau_m)\Delta t + \sqrt{\sigma^2(\tilde{x}_m; \theta)\Delta t}\, G, \tag{2.99}$$

where G is a standard Gaussian, and

$$\tilde{\mu}(\tilde{x},\tau) = \frac{\tilde{x}_t - \tilde{x}}{t-\tau}. \tag{2.100}$$

The connection between the bridge process $\tilde{X}$ and another process

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \tag{2.101}$$

is given by the Radon-Nikodym derivative:

$$d\rho_t = \rho_t \frac{\mu(x) - \tilde{\mu}(x)}{\sigma(x)} d\tilde{W}_t, \quad \rho_0 = 1. \tag{2.102}$$

By Girsanov's theorem, we can obtain the transition probability from 0 to T as

$$p(x_T, T; x_0, 0) = \int p(x_T, T; x_\tau, \tau_{M-1}\rho_{M-1}(x_\tau) dQ_{M-1}(x_\tau), \tag{2.103}$$

where Q_{M-1} is the probability measure from the Brownian bridge. In other words, constructing a Brownian bridge to connect two points can help us to obtain transition probabilities for another process, as long as we change measure appropriately. The integral in (2.103) can be easily calculated using, e.g. Monte Carlo methods.

Example 2.41. *Ornstein-Uhlenbeck process*. One of the simplest continuous-time processes is the Ornstein-Uhlenbeck process. It is similar to the Cox-Ingersoll-Ross process considered earlier in that it is mean-reverting, and only differs in the diffusion term. The process satisfies the following SDE:

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t, \quad X_0 = x_0. \tag{2.104}$$

It is a nice process in the sense that its transition density may be obtained on closed form:

$$\begin{aligned} p(X_T|x_t) = &\frac{1}{\sqrt{2\pi\frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta(T-t)}\right)}} \\ &\times \exp\left\{ -\frac{\theta\left(x_T - x_t e^{-\theta(T-t)} - \mu(1 - e^{-\theta(T-t)}\right)^2}{\sigma^2\left((1 - e^{-2\theta(T-t)}\right)} \right\}. \end{aligned} \tag{2.105}$$

Using (2.96), a tractable SDE for the conditioned Ornstein-Uhlenbeck process can be obtained as

$$\begin{aligned} dX_t = &\left[\theta(\mu - X_t) + \frac{2\theta e^{-\theta(T-t)}}{1 - e^{-2\theta(T-t)}} \left(x_T - x_t e^{-\theta(T-t)} - \mu\left(1 - e^{-\theta(T-t)}\right)\right)\right] dt \\ &+ \sigma dW_t. \end{aligned} \tag{2.106}$$

2.2.6 Girsanov measure transformations

When the formal definition of probability was set up in Definition 2.13, the need for *a* probability measure was stated. It may at times be convenient to work with several different probability measures. The necessity for some structure with which one may

compare and go between different probability measures should hereby be obvious. This section is devoted to the brief description of such a machinery. We begin by stating, without proof, a very important result regarding the relationship between two measures, whereof one is absolutely continuous with regards to the other: the Radon-Nikodym theorem.

Theorem 2.42. Radon-Nikodym's theorem. *Let $(\Omega, \mathscr{F}, \mu)$ be a finite measurable space and let ν be a finite measure on $(X, \mathscr{F})$, such that $\nu \ll \mu$. Then there exists a positive function $f : \Omega \to \mathbb{R}$ which satisfies*

i) f *is* $\mathscr{F}$*-measurable*

ii) $\int_\Omega f(x)d\mu(x) < \infty$

iii) $\nu(E) = \int_E f(x)d\mu(x), \quad$ for all Borel sets $E \in \mathscr{F}$.

The function f is called the Radon-Nikodym derivative *of ν with respect to μ on $\mathscr{F}$. It is uniquely determined almost everywhere and we write*

$$f = \frac{d\nu}{d\mu}, \quad \text{or} \quad d\nu(x) = f(x)d\mu(x). \tag{2.107}$$

(Theorem 8.11, E. Lindström, Madsen and Nielsen 2015)

Turning to a probabilistic setting, one may, given the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ augmented by the filtration $\mathscr{F}(t)$ on the time interval $[0, T]$, construct a new measure $\mathbb{Q}$ by

$$d\mathbb{Q} = L_T d\mathbb{P}, \tag{2.108}$$

where L_T is a non-negative $\mathscr{F}(T)$-measurable random variable, and T is some fixed time. If $\mathbb{E}^{\mathbb{P}}[L_T] = 1$, then $\mathbb{Q}$ is a new probability measure. With $\mathbb{P}_t$ and $\mathbb{Q}_t$ as restrictions on $\mathbb{P}$ and $\mathbb{Q}$ such that one only has knowledge of the probability measures until time t – obtaining full knowledge as $t \to \infty$ – we have that $\mathbb{Q}_t$ is absolutely continuous with respect to $\mathbb{P}_t$ for all t, and by Theorem 2.42, there exists a stochastic process

$$L_t = \frac{d\mathbb{Q}_t}{d\mathbb{P}_t}, \quad 0 \leq t \leq T. \tag{2.109}$$

Furthermore, L_t is an $(\mathscr{F}(t), \mathbb{P})$-martingale. The expectations under the different measures are connected in the following way.

Theorem 2.43. *Let the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a stochastic variable X be given, such that X has a finite first moment under $\mathbb{P}$. Let $\mathbb{Q}$ be another probability measure on $(X, \mathscr{F})$, absolutely continuous w.r.t. $\mathbb{P}$. The Radon-Nikodym derivative is then given by*

$$L = \frac{d\mathbb{Q}}{d\mathbb{P}}. \tag{2.110}$$

Assume further that $\mathscr{G} \subseteq \mathscr{F}$ is a σ-algebra. Then

$$\mathbb{E}^{\mathbb{Q}}[X|\mathscr{G}] = \frac{\mathbb{E}^{\mathbb{P}}[LX|\mathscr{G}]}{\mathbb{E}^{\mathbb{P}}[L|\mathscr{G}]} \quad \mathbb{Q} - \text{almost surely}. \tag{2.111}$$

(E. Lindström, Madsen and Nielsen 2015)

So, one can now see that it is possible to introduce new, absolutely continuous probability measures, but it should be intuitively clear that measure transformations also change the properties of the driving Wiener process and the infinitesimal dynamics of the stochastic differential equation in question. It is therefore natural – and necessary – to raise the question of how the Wiener process and the SDE change. The actual switching between probability measures entails some rather heavy probabilistic machinery.

Assume that the probability space under consideration has a (scalar) P-Wiener process W^P, and that we have defined a new probability measure $\mathbb{Q}$ for some fixed T by $d\mathbb{Q} = L_T \, d\mathbb{P}$ on $\mathscr{F}_T$. By Theorem 2.42, this will generate a process L_t on $\mathscr{F}_t$. L_t will be a non-negative martingale, and it is therefore natural to define it as the solution to the following SDE:

$$\begin{aligned} dL_t &= \varphi_t L_t dW_t^P, \\ L_0 &= 1, \end{aligned} \tag{2.112}$$

for some process φ_t. Choosing an arbitrary adapted process φ and defining the new measure as $d\mathbb{Q} = L_t d\mathbb{P}$ on $\mathscr{F}_t$ for all $t \in [0, T]$, should allow one to generate a large class of natural measure transformations from $\mathbb{P}$ to $\mathbb{Q}$.

By the Itō formula, one may easily obtain that

$$L_t = \exp\left\{ \int_0^t \varphi_s dW_s^P - \frac{1}{2} \int_0^t \varphi_s^2 ds \right\}, \tag{2.113}$$

whence one sees that it is non-negative. A φ that is sufficiently integrable ensures further that L is a martingale, and the condition $L_0 = 1$ guarantees that $\mathbb{E}^{\mathbb{P}}[L_t] = 1$, so $\mathbb{Q}$ is a new probability measure.

Heuristically, one may think of the drift of a process X_t as the expectation of dX_t given $\mathscr{F}_t$, and the squared diffusion as the expectation of $(dX_t)^2$, where dX_t is informally thought of as $X_{t+dt} - X_t$. Regarding the P-Wiener process, the drift is zero and the diffusion is one, and one can obtain

$$\mathbb{E}^{\mathbb{Q}}[dW_t^P] = \frac{\mathbb{E}^{\mathbb{P}}[L_{t+dt} dW_t^P | \mathscr{F}_t]}{\mathbb{E}^{\mathbb{P}}[L_{t+dt} | \mathscr{F}_t]} = \frac{\mathbb{E}^{\mathbb{P}}[L_t dW_t^P | \mathscr{F}_t]}{L_t} + \frac{\mathbb{E}^{\mathbb{P}}[dL_t dW_t^P | \mathscr{F}_t]}{L_t}, \tag{2.114}$$

where the first term vanishes since $L_t \in \mathscr{F}_t$, and the second term yields just $\varphi_t dt$ since $dL_t = L_t \varphi_t dW_t^P$, and $L_t \varphi_t \in \mathscr{F}_t$. It is also easy to see that the quadratic variation under $\mathbb{Q}$ is just dt, since $(dW_t^P)^2 = dt$. Thus, it is reasonable to say that the Wiener process under $\mathbb{Q}$ receives an additional drift φ, which is referred to as the *Girsanov kernel*, whereas the diffusion remains unchanged. This can be put into rigorous mathematical terms, yielding the following important result, the proof of which is outside the scope of this exposition.

Theorem 2.44. The Girsanov theorem. *Let W^P be a d-dimensional standard $\mathbb{P}$-Wiener process on $(\Omega, \mathscr{F}, \mathbb{P}, \{\mathscr{F}_t\}_{r \geq 0})$ and let φ be any d-dimensional adapted column vector process. Choose a fixed T and define the process L on $[0, T]$ by*

$$\begin{aligned} dL_t &= \varphi_t^* L_t dWt^P, \\ L_0 &= 1, \end{aligned} \tag{2.115}$$

i.e.

$$L_t = \exp\left\{ \int_0^t \varphi_s^* dW_s^P - \frac{1}{2} \int_0^t ||\varphi_s||^2 ds \right\}. \tag{2.116}$$

Assume that

$$\mathbb{E}^{\mathbb{P}}[L_T] = 1, \tag{2.117}$$

and define the new probability measure $\mathbb{Q}$ on $\mathscr{F}_T$ by

$$L_T = \frac{d\mathbb{Q}}{d\mathbb{P}}, \quad \text{on } \mathscr{F}_T. \tag{2.118}$$

Then

$$dW_t^P = \varphi_t dt + dW_t^Q, \tag{2.119}$$

where W^Q is a $\mathbb{Q}$-Wiener process. (Theorem 11.3, Björk 2009)

Regarding the requirements on $\mathbb{Q}$, Novikov's condition is useful to bear in mind.

Lemma 2.45. Novikov's condition. *If the Girsanov kernel φ is such that*

$$\mathbb{E}^{\mathbb{P}}\left[\exp\left\{\frac{1}{2}\int_0^T ||\varphi_t||^2 dt\right\}\right] < \infty, \tag{2.120}$$

then L is a martingale and $\mathbb{E}^{\mathbb{P}}[L_T] = 1$.

If one considers a more general Itō process X, a Girsanov transformation gives an addition to the drift in the form of the Girsanov kernel multiplied by the diffusion, i.e. $\mu^* = (\mu + \sigma_t\varphi_t)dt$. (Björk 2009)

The likelihood process L_t that one obtains in Radon-Nikodym's theorem is not just an abstract and practically useless entity that we define out of mathematical necessity; it can be used for Maximum Likelihood estimation in continuous models. When looking at a model, one may regard each admissible parameter θ as related to its own measure. Instead of seeing a measure and its parameter as fixed, where one has observed but one of many processes X, one can say that there are many measures, one for every admissible parameter, and equally many Wiener processes. Furthermore, the observed process X is the only one, and the problem becomes one of determining the appropriate measure.

In a more mathematical setting, one fixes the probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where $X(t)$ is a Wiener process under $\mathbb{P}$. For each parameter θ in some admissible parameter set Θ, one defines the measure transformation

$$\begin{aligned} dL_\theta(t) &= \theta L_\theta(t) dX(t), \\ L_\theta(0) &= 1, \end{aligned} \tag{2.121}$$

and the measure $\mathbb{P}_\theta$ by the Radon-Nikodym derivative as $d\mathbb{P}_\theta = L_\theta(t)d\mathbb{P}$ on the natural filtration $\mathscr{F}^X(t)$ generated by $X(t)$ until time t. One then seeks to maximise the quantity $L_\theta(t)$ with respect to θ, the interpretation of which is that the solution $\hat{\theta}$ is the most likely parameter given the observed process, or compactly, on a mathematical form,

$$\langle [\mathbb{P}_\theta]_{\theta\in\Theta\subseteq\mathbb{R}}, \Omega, \mathscr{F}, X\rangle. \tag{2.122}$$

(E. Lindström, Madsen and Nielsen 2015)

2.2.7 Discrete time approximations

A very important tool when working with stochastic differential equations is the discrete time approximation of SDEs. These discretisation schemes are built upon the stochastic Taylor expansion, which can be considered a stochastic counterpart of the regular and so well-used Taylor expansion, relying on the iterated application of Itō's formula.

Assume that the drift μ and the diffusion σ are smooth 'enough' around $X(t_0)$ and regard for $t \in [t_0, t]$ the integral form of a (unidimensional) stochastic differential equation,

$$X_t = X_{t_0} + \int_{t_0}^{t} \mu(X_s)ds + \int_{t_0}^{t} \sigma(X_s)dW_s. \tag{2.123}$$

Assume furthermore that the drift and the diffusion are time homogeneous. Applying Itō's formula to μ and σ yields

$$X_t = X_{t_0} + \mu(X_{t_0}) \int_{t_0}^{t} ds + \sigma(X_{t_0}) \int_{t_0}^{t} dW_s + R, \tag{2.124}$$

where the remainder R is given by

$$\begin{aligned} R = &\int_{t_0}^{t}\int_{t_0}^{s} \mathscr{L}^0\mu(X_\tau)d\tau ds + \int_{t_0}^{t}\int_{t_0}^{s} \mathscr{L}^1\mu(X_\tau)dW_\tau ds \\ &+ \int_{t_0}^{t}\int_{t_0}^{s} \mathscr{L}^0\sigma(X_\tau)d\tau dW_s + \int_{t_0}^{t}\int_{t_0}^{s} \mathscr{L}^1\sigma(X_\tau)dW_\tau dW_s, \end{aligned} \tag{2.125}$$

wherein

$$\mathscr{L}^0 = \mu\frac{\partial}{\partial X} + \frac{\sigma^2}{2}\frac{\partial^2}{\partial X^2}, \tag{2.126}$$

$$\mathscr{L}^1 = \sigma\frac{\partial}{\partial X}, \tag{2.127}$$

are operators. The deterministic and stochastic integrals in (2.124) are easy to find, quite simply yielding $t-t_0$ and a stochastic variable with Gaussian distribution, $N(0, t-t_0)$, respectively.

The simplest, though still very useful, scheme for discretising stochastic differential equations is obtained from using the first two terms of the stochastic Taylor expansion, i.e. using (2.124), but omitting the remainder term. This discretisation scheme is known as the *Euler-Maruyama scheme* and takes the form

$$X_{k+1} = X_k + \mu(X_k)\Delta t + \sigma(X_k)\Delta W, \tag{2.128}$$

where $\Delta t = t_{k+1} - t_k$ is the length of the time interval between the points of the discretisation, and $\Delta W = W_{k+1} - W_k$ is the corresponding increment of the Wiener process W, the distribution of which is $N(0, \Delta t)$. (E. Lindström, Madsen and Nielsen 2015)

The retention of one additional term from the remainder yields the more accurate, but less tractable, *Milstein scheme*. However, this scheme does not yield a Gaussian random variable like the Euler-Maruyama scheme, but a sum of a Gaussian and a χ^2-distributed stochastic variable, making it more difficult to handle.

Here only a very brief explanation has been given of the manner in which one may discretise stochastic differential equations. A far more thorough exposition can be found in Kloeden and Platen 1992.

Chapter 3

Algorithms for simulating diffusion bridges

3.1 Pedersen: A first approach

Pedersen 1995 looks at maximum likelihood estimation for stochastic differential equation models, which are continuous, based on discrete observations. For maximum likelihood estimation, one needs the transition probabilities of one's process in order to construct the appropriate (log-)likelihood function. However, transition probabilities for stochastic differential equations are generally difficult to handle. Although they may be obtained analytically for some (simple) models such as the Ornstein-Uhlenbeck process and the Cox-Ingersoll-Ross model, they remain intractable for many processes of interest.

In order to circumvent the problem of not knowing the continuous transition densities, it is often convenient to approximate a diffusion process by a discrete model, often, due to its simplicity, using the Euler-Maruyama method (see section 2.2.7). Hence, one obtains Discrete Maximum Likelihood (DML) estimates. However, these estimates have the unfortunate quality of being heavily biased unless the (maximum, in the case of non-equidistant observations) distance between data points is 'small'.

Under appropriate assumptions of existence and uniqueness, and the positive definiteness of $\Gamma = \sigma\sigma^*$, Pedersen proposes the introduction of M intermediate points, $t_k = \tau_0 < \tau_1 < \ldots < \tau_{M-1} < \tau_M = t_{k+1}$, bridging the gap between consecutive data points. But, as the imputed points have not actually been observed, they need to be integrated out. So, by the Markov property and the Chapman-Kolmogorov equation, an approximate transition density is given by

$$p^{(M)}(x_{k+1}|x_k) = \int \prod_{m=1}^{M} p(x_m|x_{m-1}) dx_1 \cdots dx_{M-1} = \mathbb{E}[p(x_M|x_{M-1})|x_0], \quad (3.1)$$

where $x_m = x(\tau_m)$. Thereby, one can approximate the transition densities with the Monte Carlo method, yielding

$$p^{(M)}(x_{k+1}|x_k) \approx \frac{1}{K} \sum_{k=1}^{K} p(x_M|\xi_{M-1,k}), \quad (3.2)$$

where $\xi_{M-1,k}$ are draws from $\xi_{M-1,k} \sim p(x_{M-1}|x_0)dx_{M-1}$. Pedersen then goes on to prove that $p^{(M)}(x_{k+1}|x_k) \to p(x_{k+1}|x_k)$ as $M \to \infty$ and $K \to \infty$.

This method uses the naïve dynamics of the process to construct proposals and is easy to implement but often inefficient from a computational point of view, as it essentially pays little attention to the information provided by the end-point. Instead, a barrage of skeleton paths is fired off in the general direction given by the drift, hoping that a sufficient number of them veer off their course and find their way to the end-point. Of course, this hope may be, and often proves to be, misplaced.

It is noted in Papaspiliopoulos, G. O. Roberts and Stramer 2013 that simulating paths forward in this way and accepting them if they reach the desired end-point $X_T = x_T$ renders exact draws from the desired process. The probability of reaching the exact end-point is, of course, zero, so by the Markov property of diffusions and Bayes' theorem, one introduces a weight to each path as the density of the end-point, which can be approximated by the Euler-Maruyama method provided that the subdivision of the interval is small enough. Unfortunately, as the subdivision becomes finer, and the bias from the Euler-Maruyama approximation decreases, the weights approach zero (a.s.), and a single path is assigned a weight much larger than the others, that is, placing all one's proverbial eggs in one basket.

For the above mentioned reasons, a great deal of work has been put into attempts to create more efficient simulation algorithms.

3.2 Durham-Gallant: A simple but effective algorithm

A significant improvement over the aforementioned, naïve algorithm was introduced in Durham and Gallant 2002, where they regard the approximation of the transition density as a Monte Carlo estimate of an expectation, and applying thereon an importance sampler, which, based on the linearised dynamics of the process, is close to optimal.

Durham and Gallant propose an importance sampler that considers x_m and x_M to be fixed and then draws x_{m+1} from

$$\begin{aligned} p(x_{m+1}|x_m, x_M) &= \frac{p(x_{m+1}|x_m)p(x_M|x_{m+1})}{p(x_M|x_m)} \\ &\approx \phi(x_{m+1}; x_m + \mu(x_m)\Delta t, \sigma^2(x_m)\Delta t) \\ &\cdot \phi(x_M; x_{m+1} + \mu(x_m)\tilde{\Delta} t, \sigma^2(x_m)\tilde{\Delta} t) \\ &/\phi(x_M; x_m + \mu(x_m)\bar{\Delta} t, \sigma^2(x_m)\bar{\Delta} t), \end{aligned} \tag{3.3}$$

where $\phi(\cdot; a, B)$ denotes the density of a multivariate Gaussian with mean a and covariance B, $\tilde{\Delta} t = \tau_M - \tau_{m+1}$ and $\bar{\Delta} t = \tau_M - \tau_m$. Some manipulation of the resulting Gaussian yields the *Modified diffusion bridge* (MDB).

$$q(x_{m+1}|x_m, x_M) = \phi\left(x_{m+1};\ x_m + \frac{x_M - x_m}{\tau_M - \tau_m}\Delta t, \frac{M-m-1}{M-m}\sigma^2(x_m)\Delta t\right), \tag{3.4}$$

where q denotes the proposal density. It should be noted that this sampler turns out to be almost identical to a Brownian bridge, differing only by the factor $(M-m-1)/(M-m)$ in the variance.

Durham and Gallant's approach can be likened with simulating from a stochastic process conditioned on reaching the end-point. Noting that the diffusion is the same in the process conditioned to hit the end-point as in the unconditioned process (see section

2.2.5), and approximating the drift and diffusion of the SDE by the constant $\mu(T, x_T; \theta)$ and $\sigma(T, x_T; \theta)$, the unconditional process becomes a regular, scaled Brownian motion with drift, the bridge process for which is a Brownian bridge, which is not difficult to handle (see Example 2.40).

Improving upon Pedersen's simulation method, one may propose paths based on a stochastic process that reach the end-point at time T almost surely. In Delyon and Hu 2006, the process in question is suggested to take the form

$$dX_t = \frac{x_T - X_t}{T - t}dt + \sigma(t, X_t; \theta)dW_t, \tag{3.5}$$

which is a process with a drift that depends on x_T, *not* a conditioned process. Delyon and Hu go on to show that the distribution of the original, target process is absolutely continuous with regards to the distribution of the solution of this new SDE, (3.5), affirming the validity of the method. Using the Euler-Maruyama approximation leads to the discrete version,

$$X_{k+1} = X_k + \frac{x_T - X_k}{T - \tau_k}dt + \sigma(\tau_k, X_k; \theta)\sqrt{\frac{T - \tau_{k+1}}{T - \tau_k}}dW_k, \tag{3.6}$$

which, quite plainly, is Durham and Gallant's MDB construct.

One of the drawbacks of the modified diffusion bridge is that it is only efficient when the drift can be said to be approximately constant over the time interval in question. If the process exhibits strong non-linearities, the MDB is likely to yield poor results.

3.3 Lindström: A mixture

Whilst Durham and Gallant's modified diffusion bridge is simple, reasonably effective and commonly used, it does have a great fault in that it provides very poor results for diffusions whose dynamics is dominated by the drift rather that the random diffusion. The sampler is, in fact, entirely independent of the drift of the process, effectively ignoring some of the dynamics of the process. In his 2012 paper, Erik Lindström seeks to ameliorate this problem by combining the MDB with the naïve sampler. Starting from the realisation that it is preferable to have a proposal distribution that is too heavy-tailed than too light-tailed, the mean squared error (MSE) is studied. It is known that

$$\mathrm{MSE}[x_{t+\Delta t}|x_t] = \mathbb{V}[x_{t+\Delta t}|x_t] + (\mathrm{Bias}[x_{t+\Delta t}|x_t])^2, \tag{3.7}$$

and that for the Euler-Maruyama scheme

$$\mathbb{V}[x_{t+\Delta t}|x_t] \approx \sigma(t, x_t)\sigma(t, x_t)^T\Delta t, \tag{3.8}$$

$$\mathrm{Bias}[x_{t+\Delta t}|x_t] \approx c\Delta t, \tag{3.9}$$

where c is some column vector.

Lindström first suggests the addition of the conditional mean square errors to their respective densities in the numerator of (3.3), which, with the variances of the first and second densities in the numerator of (3.3) denoted P and Q, respectively, yields

$$\tilde{P} = P + cc^T(\tau_{m+1} - \tau_m)^2, \tag{3.10}$$

$$\tilde{Q} = Q + cc^T(\tau_M - \tau_{m+1})^2. \tag{3.11}$$

However, Lindström notes, it is more robust to only augment the latter, yielding

$$\hat{P} = P, \tag{3.12}$$

$$\hat{Q} = Q + cc^T(\tau_M - \tau_{m+1})^2. \tag{3.13}$$

This sampler, henceforth referred to as the first order Lindström bridge, will have both the naïve sampler and the MDB as special cases, with $c = \infty$ yielding the naïve sampler and $c = 0$ the MDB.

It is also possible to include higher order terms in the bias, hence obtaining the second order Lindström bridge as

$$\bar{P} = P, \tag{3.14}$$

$$\bar{Q} = Q + \gamma_2(\tau_M - \tau_{m+1})^2 + \gamma_4(\tau_M - \tau_{m+1})^4. \tag{3.15}$$

Due to the high computational cost of reducing bias, the spacing between consecutive data points is probably chosen such that some bias remains, although the variance dominates. As a first approximation one may say that the squared bias is a fraction of the variance, approximating the MSE by

$$\mathrm{MSE}[x_{m+1}|x_m] \approx (1+\alpha)P = P + cc^T(\tau_{m+1} - \tau_m)^2, \tag{3.16}$$

where α is the fraction, which must be chosen by the user. Hence, a heuristic expression for cc^T can be obtained,

$$cc^T_{Heur} = \alpha \frac{P}{(\tau_{m+1} - \tau_m)^2}. \tag{3.17}$$

A similar argument can be given for the heuristic obtention of expressions for γ_2 and γ_4 for the second order Lindström bridge. Here, γ_2 is chosen to be cc^T_{Heur} from (3.17), and γ_4 is chosen such that the first term of the bias dominates the second. This amounts to solving

$$\gamma_4(\tau_{m+1} - \tau_m)^4 = \varepsilon\gamma_2(\tau_{m+1} - \tau_m)^2, \tag{3.18}$$

where ε is some small constant, yielding $\gamma_4 = \varepsilon\alpha\frac{P}{(\tau_{m+1}-\tau_m)^4}$, and $\gamma_2 = \alpha\frac{P}{(\tau_{m+1}-\tau_m)^2}$. In both cases, less emphasis is placed on coming observation when the time until the end, $\tau_M - \tau_{m+1}$, is large compared with the length of the individual subintervals, $\tau_{m+1} - \tau_m$. The variance of the bridge is diminished to a smaller degree with this construct than with the MDB, which may be positive, as too light tails might lead to a rate of convergence slower than $\sqrt{N}$.

Explicit expressions for the sampler may be obtained. For the first order Lindström bridge, these expressions are obtained in Eqs (3.21)–(3.22). The expressions for the second order bridge is obtained similarly.

Using the aforementioned, heuristically obtained values for cc^T, together with the Euler-Maruyama approximation leads to the following expressions for P and Q:

$$\hat{P} = \sigma(\cdot)\sigma(\cdot)^T(\tau_{m+1} - \tau_m), \tag{3.19}$$

$$\hat{Q} = \sigma(\cdot)\sigma(\cdot)^T\left(\tau_M - \tau_{m+1} + \alpha\frac{(\tau_M - \tau_{m+1})^2}{\tau_{m+1} - \tau_m}\right). \tag{3.20}$$

Hence, with $K_0 = P(P+Q)^{-1}$, and $\hat{K}_0 = \hat{P}(\hat{P}+\hat{Q})^{-1}$ being the Kalman gains for the MDB, and the first order Lindström bridge, respectively, the conditional mean and variance are given as

$$\mu_{LB}(x_m) = (I - \hat{K}_0K_0^{-1})a + \hat{K}_0K_0^{-1}(a + K_0(x_M - a - b)), \tag{3.21}$$

where a and b denotes the expectations of the first and second densities in the numerator of (3.3), and

$$\Psi_{LB}(x_m) = (I - \hat{K}_0 K_0^{-1})P + \hat{K}_0 K_0^{-1}(I - K_0)P, \tag{3.22}$$

respectively. The quantity

$$\hat{K}_0 K_0^{-1} = I \frac{(\tau_{m+1} - \tau_m)(\tau_M - \tau_m)}{(\tau_{m+1} - \tau_m)(\tau_M - \tau_m) + \alpha(\tau_M - \tau_{m+1})^2} \tag{3.23}$$

acts as a weight, combining the naïve Pedersen sampler with the MDB. The proposal density for the Lindström bridge is thus

$$q(x_{m+1}|x_m, x_M) = N\big(x_{m+1}; x_m + \mu_{LB}(x_m), \Psi_{LB}(x_m)\big). \tag{3.24}$$

Note, however, that this construct requires a tuning parameter α, which might be a drawback.

3.4 Whitaker et al: Solutions from ordinary differential equations

Whitaker et al. 2016 propose to handle non-linearities in the drift by partitioning the process into a deterministic function ξ_t and a residual stochastic process, R_t, upon which Durham and Gallant's modified diffusion bridge can be applied. Thus, $X_t = \xi_t + R_t$, satisfying

$$d\xi_t = f(\xi_t)dt, \quad \xi_0 = x_0, \tag{3.25}$$

$$dR_t = [\mu(X_t) - f(\xi_t)]dt + \sigma(X_t)dW_t, \quad R_0 = 0. \tag{3.26}$$

All that remains is to find an appropriate process ξ_t. Whitaker et al. propose two different possibilities for ξ_t, namely, a mere subtraction of the drift, and a more advanced subtraction of the linear noise approximation of the process.

The first approach is to take $\xi_t = \eta_t$ and $f(\cdot) = \mu(\cdot)$, yielding

$$d\eta_t = \mu(\eta_t)dt, \quad \eta_0 = x0, \tag{3.27}$$

$$dR_t = [\mu(X_t) - \mu(\eta_t)]dt + \sigma(X_t)dW_t, \quad R_0 = 0. \tag{3.28}$$

Then, the application of the modified diffusion bridge on the residual process yields the desired sampler. Finding η_t is merely an exercise in solving ordinary differential equations, so that should not pose a problem. However, the sufficient linearity of the residual process is but a hope, and it is quite possible that this subtraction does not account for enough of the non-linearity to yield satisfactory results.

For the above mentioned reason, Whitaker et al. propose the additional subtraction of an approximation of the conditional expectation, obtained from the linear noise approximation. Offering only a very brief account of the linear noise approximation, the interested reader is referred to e.g. Fearnhead, Giagos and Sherlock 2014.

A Taylor expansion of $\mu(\cdot)$ and $\sigma(\cdot)$ around η_t, keeping the first and the second terms of the expansion $\mu(\cdot)$ and only the first term of the expansion of $\sigma(\cdot)$ (assuming that the random term is 'small'), yields

$$d\hat{R}_t = J(\eta_t)\hat{R}_t dt + \sigma(\eta_t)dW_t, \tag{3.29}$$

where $J(\eta_t)$ is the Jacobian matrix of $\mu(\eta_t)$. Hence, one finds that $\hat{R}_t|\hat{R}_0 = \hat{r}_0 \sim N(m_t\hat{r}_0,\, m_t\Phi_t m_t^T)$, where m_t and Φ_t solves the following system of ODEs:

$$\frac{dm_t}{dt} = J(\eta_t)m_t, \quad m_0 = I_d, \tag{3.30}$$

$$\frac{d\Phi_t}{dt} = m_t^{-1}\Gamma(\eta_t)m_t^{-T}, \quad \Phi_0 = 0, \tag{3.31}$$

where I_d is the d-dimensional identity matrix, and $\Gamma = \sigma\sigma^T$. An approximation of the conditional expectation $\rho_t = \mathbb{E}[R_t|r_0, x_T]$ is thereby obtained as

$$\hat{\rho}_t = \mathbb{E}[\hat{R}|r_0, x_T] = m_t\Phi_t m_T^T(m_T\Phi_T m_T^T)^{-1}(x_T - \eta_T). \tag{3.32}$$

Once again, any difficulty in obtaining the above quantities is easily overcome; the solution of the system of ODEs, (3.27), (3.30) and (3.31), is straightforward to obtain. This construct should also capture more of the non-linear behaviour of the drift, thereby offering better results than the first approach.

Chapter 4

An adaptive algorithm

Even the best simulation methods described in Chapter 3 may yield unsatisfactory results in cases where the the expected path of the process, ρ_t, exhibits very significant non-linear behaviour. The last simulation method described in Section 3.4, based on the linear noise approximation, should yield the best results, but it still relies on the creation of an approximate expected path that is ultimately based on the corresponding ordinary differential equation, with apposite corrections.

The new method proposed herein is based on the methods of Whitaker et al. 2016 – more specifically, the linear noise approximation – and uses the repetitive nature of the Metropolis-Hastings framework, in which the simulations are carried out, to find the expectation. Each iteration provides a little information about the conditioned process in question, and Adaptive Markov chain Monte Carlo (see e.g. Andrieu and Thoms 2008) is used to update the approximation of the expected path, improving the approximation as more simulations are made available.

Whitaker et al. base their algorithm on the partition of the process into a deterministic function ξ_t^*, which accounts for the drift of the process, and the residual stochastic process R_t, which accounts for the randomness,

$$X_t = \xi_t^* + R_t, \tag{4.1}$$

and seek to find good approximations, ξ_t, of the optimal function ξ_t^*. The efficiency of the simulations depends on this choice. The same route is taken here, which naturally leads to the question of how to find this deterministic function upon which so much hinges. Ideally, one would want to find the bridge, ξ_t^*, that minimises the variance of the Importance sampling approximation, i.e. one wants to find the optimal bridge

$$\xi_t^* = \min_{\xi_t} \mathbb{V}[r_\xi(X_t)]. \tag{4.2}$$

However, it is far from simple to find such an optimal bridge. Looking at the SDE of the exact bridge from Section 2.2.5 and applying an Euler-Maruyama discretisation, one obtains something that is conditionally Gaussian. With a residual process, R_t, in the form of a Brownian Bridge, and thereby with zero expectation, the task of guiding the process in whatever direction it should go is borne entirely by the (optimal) deterministic function ξ_t^*. Hereby, the optimal function ξ_t^{Opt} is obtained as the expected path of the bridge process, i.e.

$$\xi_t^{Opt} = \mathbb{E}[X_t | x_T, x_0]. \tag{4.3}$$

Now, this expectation is normally not easy to obtain, cf. Section 2.2.5, at least not analytically. It may, however, be obtained numerically using the Monte Carlo method.

The transition density of the bridge process is often of great interest but rarely tractable. To bridge the gap, additional data points are imputed into the interval over which the bridge spans, but since the imputed points are unknown, one has to integrate them out, yielding

$$\begin{aligned} p(x_{(0,T]}|x_0, x_T, \theta) &= \int \prod_{m=0}^{M} p(x_{m+1}|x_m, \theta) dx_{1:M-1} \\ &= \int \frac{\prod_{m=0}^{M} p(x_{m+1}|x_m, \theta)}{q(x_{m+1}|x_m, x_M, \theta)} q(x_{m+1}|x_m, x_M, \theta) dx_{1:M-1} \\ &= \int \frac{\tilde{p}(x_{0:M})}{q(x_{0:M})}\, q(x_{0:M})\, dx_{1:m-1} \\ &= \mathbb{E}_q[r(X)] = p, \end{aligned} \tag{4.4}$$

where the last two lines only introduce convenient notation. Thus,

$$\mathbb{V}[r(X)] = \mathbb{E}[r(X)^2] - (\mathbb{E}[r(X)])^2 = \mathbb{E}[r(X)^2] - p^2, \tag{4.5}$$

and, focusing on the first term and multiplying by q_{Opt}/q_{Opt}, i.e. 1,

$$\mathbb{E}[r(X)^2] = \int \frac{\tilde{p}(x_{0:M})^2}{q(x_{0:M})^2} \frac{q_{Opt}(x_{0:M})^2}{q_{Opt}(x_{0:M})^2}\, q(x_{0:M})\, dx_{1:m-1}, \tag{4.6}$$

which, setting a proposed 'optimal' proposal density, $q_{Opt} = \tilde{p}/p$, to be the sought product of densities, normalised by the full density, yields

$$\mathbb{E}[r(X)^2] = p^2 \int \frac{q_{Opt}(x_{0:M})^2}{q(x_{0:M})^2}\, q(x_{0:M})\, dx_{1:m-1}. \tag{4.7}$$

Hence, by Jensen's inequality and as the densities are positive, the variance is zero when $q = q_{Opt}$, i.e. the optimal proposal density is q_{Opt}.

As proposed in Whitaker et al. 2016, the process is partitioned into one deterministic and one random part. The modified diffusion bridge is then applied to a residual process obtained from subtracting an approximation of the expected path of the process. The simulations themselves are then carried out in a Metropolis-Hastings framework.

Mirroring the argument in Delyon and Hu 2006, the use of the modified diffusion bridge ensures a connection between the continuous-time, conditioned process and the simulated process, see Section 3.2.

From this argument, it is proposed that already accepted realisations of the process are used to adaptively update the expected path whence future realisations are created. Thereby, the expected path that forms the basis for simulations should approach the expected path of the conditioned process of interest, cf. (4.3). In a Bayesian sense, this should improve the prior distribution of the analysis.

A first course of action might be to start from some initial path and then form the point-wise mean of every obtained realisation. The accept-reject method of simulation should ensure that the obtained Markov chain indeed has the correct distribution, and, by the law of large numbers, the mean thus converges to the expectation. A sensible choice of initial distribution is the best approximation obtained from Whitaker et al. 2016.

The convergence of the mean, however, might prove to be very slow, and, if the random part takes the realisations far away from the expected path, the mean will initially have the shape of a somewhat smoothed realisation, which need not resemble the actual expected path. If this is the case, it is reasonable to assume that convergence will be slow, as the expected path around which realisations are created is initially being distorted rather than improved by earlier realisations. In the limit this is of course not a problem, but one might not have the patience to wait for convergence, so it is desirable to discard as few realisations as possible.

It would therefore be advantageous to approximate the expected path by some other method, which is less sensitive to initial distortion by early realisations. Inspiration for such a method can be found in the theory of Fourier series and least squares regression.

Removing the linear noise approximation of the expected path, one is left with a remainder, the properties of which one wishes to capture, and thereby obtain a better approximation of the expected path. The remainder begins and ends at the correct points, but non-linearities may cause the in-between path to diverge from the true path. Since the remainder is zero at both the initial point and the end-point, one can regard it as a part of an odd, periodic function and seek to approximate it accordingly.

It is well-known from Fourier analysis that a sensible choice of basis functions for the representation of a 'periodic' function of interest is an infinite number of sinusoids. Choosing sine functions, $\sin\left(\frac{n\pi t}{T}\right)$, and cosine functions, $cos\left(\frac{n\pi t}{T}\right)$, $n \geq 1$, where T is the length of the interval in question, one has that

$$\int_0^T \sin\left(\frac{n\pi t}{T}\right)\sin\left(\frac{m\pi t}{T}\right)dt = \begin{cases} T/2, & m = n \\ 0, & m \neq n \end{cases}, \tag{4.8}$$

$$\int_0^T \cos\left(\frac{n\pi t}{T}\right)\cos\left(\frac{m\pi t}{T}\right)dt = \begin{cases} T/2, & m = n \\ 0, & m \neq n \end{cases}, \tag{4.9}$$

$$\int_0^T \sin\left(\frac{n\pi t}{T}\right)\cos\left(\frac{m\pi t}{T}\right)dt = 0, \quad \forall\, n, m \geq 1, \tag{4.10}$$

i.e. the sine and cosine functions form an orthogonal set. This set is furthermore an orthogonal basis for the Hilbert space $L^2[0, T]$.

Theorem 4.1. *Let S be a closed convex subset of a Hilbert space H. For every point $x \in H$ there exists a unique point $y \in S$ such that*

$$||x - y|| = \inf_{z \in S} ||x - z||. \tag{4.11}$$

(Theorem 3.9.2, Debnath and Mikusinski 1999)

That is to say, there is a best approximation of an element in H. One can approximate a function in L^2 by a suitable number K of sine functions. This partial Fourier series converges almost everywhere. (Carleson 1966)

Assuming then that the drift of the conditioned process, which one tries to approximate, is in $L^2[0, T]$, i.e. assuming that it is square integrable, the best approximation of the process is a partial Fourier series. Saying that it is odd implies that every cosine coefficient is zero, leaving only sine functions.

With each accepted realisation, more information is made available, and it is therefore proposed that a regression model based on the aforementioned sine functions is recursively fitted to the remainder. Summing the linear noise approximation and the

remainder approximation should yield a good approximation of the actual expected path, and thereby improve simulations.

Regression can be written on the form

$$Y_m = \Phi\theta + e_m, \tag{4.12}$$

where Y_m is the remainder of simulation m at every time t_l, e_m is a Gaussian random noise vector, θ is a parameter vector, and Φ is a matrix of basis functions given explicitly as

$$\Phi = \begin{bmatrix} \sin\left(\frac{\pi t_0}{T}\right) & \dots & \sin\left(\frac{\pi t_l}{T}\right) & \dots & \sin\left(\frac{\pi T}{T}\right) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sin\left(\frac{k\pi t_0}{T}\right) & \dots & \sin\left(\frac{k\pi t_l}{T}\right) & \dots & \sin\left(\frac{k\pi T}{T}\right) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sin\left(\frac{K\pi t_0}{T}\right) & \dots & \sin\left(\frac{K\pi t_l}{T}\right) & \dots & \sin\left(\frac{K\pi T}{T}\right) \end{bmatrix}. \tag{4.13}$$

Note that Φ does not depend on which realisation m of the simulation is regarded, and that L is the number of time steps and K the number of basis functions. An estimator of θ using M realisations is then

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{m=1}^{M} (Y_m - \Phi^T\theta)^T (Y_m - \Phi^T\theta), \tag{4.14}$$

which through differentiation can be made into a useful formula as

$$\hat{\theta}_M = \left(\sum_{m=1}^{M} \Phi\Phi^T\right)^{-1} \left(\sum_{m=1}^{M} \Phi Y_m\right). \tag{4.15}$$

Denoting $\Phi\Phi^T$ by R, and realising that $\sum_{m=1}^{M} \Phi\Phi^T$ reduces to the product MR since the values of the basis functions are fixed for all m, yields a recursive formula:

$$\hat{\theta}_M = \hat{\theta}_{M-1} + \frac{1}{M}\left(R^{-1}\Phi Y_M - \hat{\theta}_{M-1}\right). \tag{4.16}$$

Straightforward calculations show that the estimator of θ is consistent.

An estimate of the expected path is then given as the sum of the linear noise approximation and this new correction, i.e.

$$\hat{\rho_t}^{(M)} = \hat{\rho_t}^{LNA} + \Phi^T\hat{\theta}_M. \tag{4.17}$$

The approximation improves continually as more information about the process is made available with every iteration of the Metropolis-Hastings sampler. If one furthermore keeps the number of basis functions low, one introduces a resilience towards the erratic behaviour of individual realisations, the effect of which is initially large. However, as one has truncated the number of basis functions, the assumption that the (Fourier) sine series accurately represents the actual expectation function does not necessarily hold true. Thus, one has simultaneously introduced bias.

To ameliorate this problem it is proposed that one, after some burn-in time, switches to calculating the point-wise mean. This ensures that the expectation actually approaches the sought expected path. The actual change is simple. After the burn-in

period, one just updates the approximation by the regular recursive formulation of the mean,

$$\hat{\rho}_t{}^{(M)} = \hat{\rho}_t{}^{(M-1)} + \frac{1}{M}\left(X_M - \hat{\rho}_t{}^{(M-1)}\right), \tag{4.18}$$

instead of the recursive least squares update above.

Building upon example 2.41 and taking expectations, it is possible to solve the ordinary differential equation corresponding to the expected path. In figure 4.1 the analytically obtained expected path is compared to the approximation of the expected path obtained with the proposed sampler.

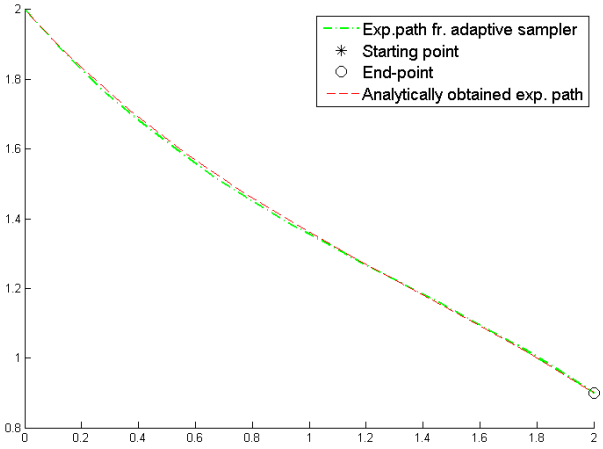


Figure 4.1: *The figure shows both the analytically obtained expected path for the Ornstein-Uhlenbeck process and the mean of the Markov chain obtained with the proposed algorithm. They are quite close to each other.*

The proposed algorithm adaptively updates the proposal density, parameterised by some $\theta \in \Theta$, thereby hoping to find the optimal one. Adaptive Markov chain Monte Carlo, however, is not an entirely unproblematic subject. More specifically, carelessness in constructing the adaptive algorithm may disturb the ergodicity of the constructed Markov chain, and even lead to the loss of the desired distribution as stationary distribution. This is, naturally, rather bad, as one may consider the fact that the chain actually has the desired distribution as stationary distribution a prerequisite for sampling from it. One may also happen to construct processes such that estimators of the type

$$\hat{I}_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(X_i), \tag{4.19}$$

which are often of great interest, become inconsistent.

Many remedies can be proposed to counter the aforementioned problems. One may, for example, stop the adaptation when one has attained a sufficiently good, albeit suboptimal, value for θ, according to some appropriate stopping criterion. Adaptive Markov chain Monte Carlo algorithms that produce samples asymptotically distributed according to the desired distribution are by no means impossible to construct. However, their construction does require some care.

An important criterion that allows the retention of the desired ergodic properties is vanishing adaptation. That is to say, the dependence of the parameter θ_i at iteration i on recent states is diminishing as i grows. (Andrieu and Thoms 2008) In the proposed algorithm, the mean is used to ensure convergence, and as the effect of the last state on θ diminishes as more samples are added to the estimate, the constructed chain should have the desired properties.

To formalise the above argument, define, for any $\theta, \bar{\theta} \in \Theta$,

$$D_{TV}(\theta, \bar{\theta}) \stackrel{\text{def.}}{=} \sup_{x \in \mathsf{X}} ||P_\theta(x, \cdot) - P_{\bar{\theta}}(x, \cdot)||_{TV}, \tag{4.20}$$

with regards to total variation, where X is the state space, and P_θ is the transition kernel of the process with parameter θ. Furthermore, define, for $x \in \mathsf{X},\ \theta \in \Theta$ and any $\varepsilon > 0$,

$$M_\varepsilon(x, \theta) \stackrel{\text{def.}}{=} \inf\{n \geq 0, ||P_\theta^n(x, \cdot) - \pi_\theta||_{TV} \leq \varepsilon\}, \tag{4.21}$$

where π_θ is the probability measure with regards to θ. Next, we state three fundamental assumptions, by which the desired results may be obtained.

A. For any $\theta \in \Theta$, P_θ is a transition kernel with a unique stationary distribution π_θ.

B. *(Vanishing adaptation)* The sequence $\{D_{TV}(\theta_n, \theta_{n-1})\}_{n \geq 1} \xrightarrow{p} 0$.

C. *(Containment)* For any $\varepsilon > 0$,

$$\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P}[M_\varepsilon(X_n, \theta_n) \geq M] = 0. \tag{4.22}$$

That is, the sequence $\{M_\varepsilon(X_n, \theta_n)\}_{n \geq 0}$ is bounded in probability.

Under the aforementioned assumptions one can show that the marginal distribution will indeed converge to the desired one and, furthermore, establish a strong law of large numbers for adaptive Markov chain Monte Carlo algorithms. (Atchadé et al. 2011)

In the case of the proposed sampler, the target distribution is essentially a multivariate Gaussian distribution, the expectation of which is adaptively improved, in a Metropolis-Hastings framework. Bai, G. O. Roberts and Rosenthal 2009 set up conditions under which an adaptive Metropolis-Hastings set-up has the necessary ergodicity property – more specifically, they address the containment condition – and go on to explicitly show that this is the case for a Gaussian target density where the variance is adaptively updated. Slight modifications of the arguments of Bai, G. O. Roberts and Rosenthal 2009 lead one to deduce that the desired ergodicity property is valid for the proposed sampler.

This proposed sampler introduces a robustness not present in the earlier samplers in Whitaker et al. 2016; the adaptive sampler manages to uphold high acceptance ratios even for extreme cases. Although interest in these extreme cases may seem to be purely academic at a first glance, these extremes may appear in applications. For example, when parameters are to be estimated in a Gibbs sampling framework it is necessary to explore the entire parameter space. It is then natural to assume that extreme cases with regards to the set of parameters currently examined will occur, lending a further raison d'être to the method herein proposed.

Chapter 5

Simulation study

The proposed sampler has been compared to the closely related samplers from Whitaker et al. 2016. Whitaker et al. show that their method is essentially as good as, or better than its predecessors, and therefore the performance of the proposed method is not compared to the antecedent methods. For the first model, the proposed sampler was compared to both the samplers in Whitaker et al. 2016, but for the second one, the results of the ODE sampler were found to be inferior to those of the LNA sampler, and they were therefore omitted.

The methods were tested by repeated simulation in a Metropolis-Hastings independence framework, i.e. new candidates do not directly depend on the previous value in the chain. In order to see that realisations are generated in the right region, and that the simulation creates viable bridges, the proposed method is evaluated with regard to efficiency, measured by the estimated acceptance ratio. This method of evaluation is also used by Whitaker et al., facilitating comparison with their methods. A bridge was simulated $10\,000$ times for each sampler and the Metropolis-Hastings acceptance ratio was estimated by the share of accepted bridges.

5.1 Cox-Ingersoll-Ross model

As a relatively simple starting model, we first regard the Cox-Ingersoll-Ross model (CIR) from Example 2.36. The CIR model is comparatively easy to work with, whilst still possessing some rather interesting, non-linear features. A multitude of runs are necessary to explore the full features of the proposed sampler, and calculations are light enough for this not to be an insurmountable task.

The parameters of the model were chosen to be $\kappa = 1$, $\theta = 5$, and $\sigma = 0.1$, with an initial value of $x_0 = 6$. The time horizon was chosen to be $T = 2$, and five sine basis functions were used. The chosen parameters ensures that the process is positive since

$$2\kappa\theta \geq \sigma^2, \tag{5.1}$$

i.e. they fulfil the Feller condition.

Fig. 5.1 shows estimated acceptance ratios for bridges to 14 different end-points at different distances from the end-point of the deterministic initial value problem ranging from -0.35 to $+0.35$ in steps of 5, with each simulation consisting of $10\,000$ iterations. Each bridge is simulated six times, with the number of imputed values ranging from 10 to 35 in steps of 5 (Figs 5.1a–5.1f). The proposed sampler consistently achieves

acceptance ratios on a par with, or better than the LNA sampler, and improves slowly as the number of imputed values increases. The samplers from Whitaker et al. 2016 perform worse for few imputed values, where the assumption of approximate linearity between imputed points breaks down, but improve rapidly as the number increases, with the LNA sampler approaching and eventually attaining the results of the proposed sampler. For the ODE sampler, the results deteriorate very quickly as the end-point moves away from the end-point of the solution to the corresponding initial value problem (IVP), an effect observed neither in the LNA, nor in the proposed sampler, or at least to a much lesser extent.

Looking more closely at a bridge to a single end-point – here chosen roughly $10\,\%$ above the IVP end-point, shown in Fig. 5.2, and simulated with 15 imputed points and five sine basis functions – one may study the evolution of the acceptance ratio as the loop progresses. Estimated acceptance ratios, obtained at each iteration from dividing the number of accepted trajectories thus far by the number of iterations, can be seen in Fig. 5.3. The samplers proposed in Whitaker et al. 2016 are constant over the entire loop, with only the uncertainty of the estimation distorting the horizontal line. Fig. 5.3 shows that the adaptive sampler performs significantly better than the other samplers.

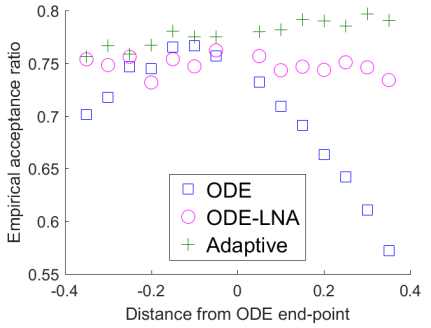
The proposed, adaptive sampler improves as the loop progresses, presumably until the subtracted expected path is undistinguishable from the actual expected path of the process. Merely taking the mean yields fairly poor results if the obtained realisations are noisy. This is attributable to a somewhat slow convergence. To speed up convergence, albeit with the addition of some bias, the approximation based on sine functions is used for the first $2\,000$ iterations, after which we take the mean to ensure proper convergence.

The deterioration of the linear noise approximation is also clearly visible in Fig. 5.2. The deterministic path that Whitaker et al. subtracts in order to form their residual process is shown in dashed magenta, and for this strongly non-linear path, the linear noise approximation differs significantly from the Monte Carlo estimate of the expected path, yielding a substantial, non-linear residual process.

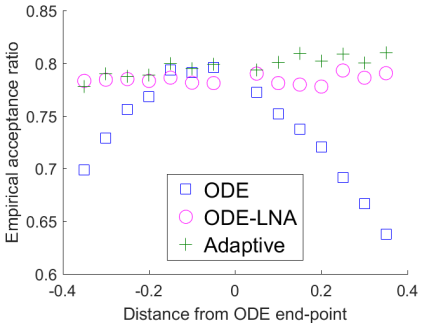
5.2 Stochastic Lorenz model

Next, we regard the stochastic Lorenz model from Example 2.38. This is a complex model owing to the chaotic behaviour of its deterministic counterpart and near periodicity. It is also three-dimensional, which adds to its difficulty.

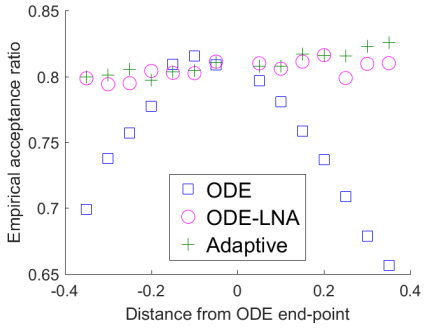
Starting point were chosen as $x_0^{\text{easy}} = [12,\ 18,\ 24]$ for an easy test case, and as $x_0^{\text{diff.}} = [-2.32,\ 0.771,\ 25.6]$ for a more difficult case. Trajectories were then simulated therefrom for appropriate time horizons of $T = 0.1$ for the easy case and $T = 0.3$ for the difficult one. In order to test the sampler on realistic bridges, the end-points were obtained by simulating $1\,000$ unconditional trajectories from the two starting points, and choosing the desired number of end-points (here 10 end-points were used) amongst end-points of the obtained, unconditional trajectories. In order to reduce the number of necessary runs, whilst still illustrating the point, end-points far from the ODE end-point were chosen with somewhat higher probability, so as not to have an abundance of data points close to the ODE end-point – bringing little new information to the table – and very few far away from it – where our main interest lies. These end-points were then used to simulate bridges from the starting point. Parameters were chosen as $s = 10$, $r = 28$, $b = 8/3$, and $\sigma = 2$, so as to obtain a model with two attractors.



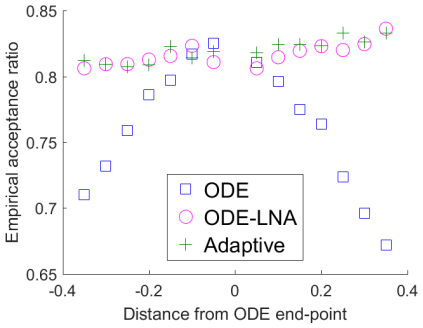
(a) *Acceptance ratios with 10 imputed values.*



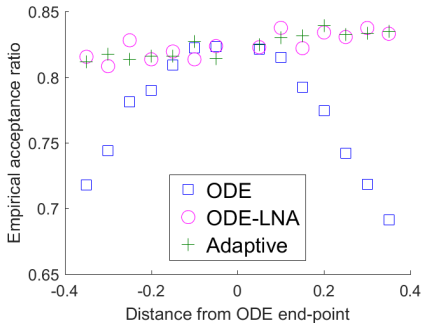
(b) *Acceptance ratios with 15 imputed values.*



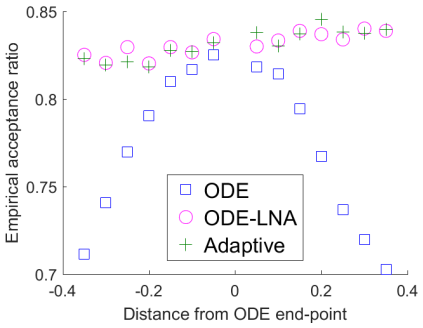
(c) *Acceptance ratios with 20 imputed values.*



(d) *Acceptance ratios with 25 imputed values.*



(e) *Acceptance ratios with 30 imputed values.*



(f) *Acceptance ratios with 35 imputed values.*

Figure 5.1: *The figures show estimated acceptance ratios for the Cox-Ingersoll-Ross model from $K = 10\,000$ iterations, for 14 end-points at different distances from the end-point of the solution of the deterministic initial value problem. Values for the proposed, adaptive sampler, and the ODE sampler and the LNA sampler from Whitaker et al. are shown as green pluses, blue squares and magenta-coloured circles, respectively. For few imputed values, the adaptive sampler yields higher acceptance ratios than both the other samplers, and as the number of imputed points increases, the LNA sampler closes in on the adaptive sampler.*

Two starting points were considered: one 'easy', going around the outside of the attractor; and one difficult, placed near a bifurcation, so that some trajectories went into one attractor and some into the other. The solution to the ordinary differential

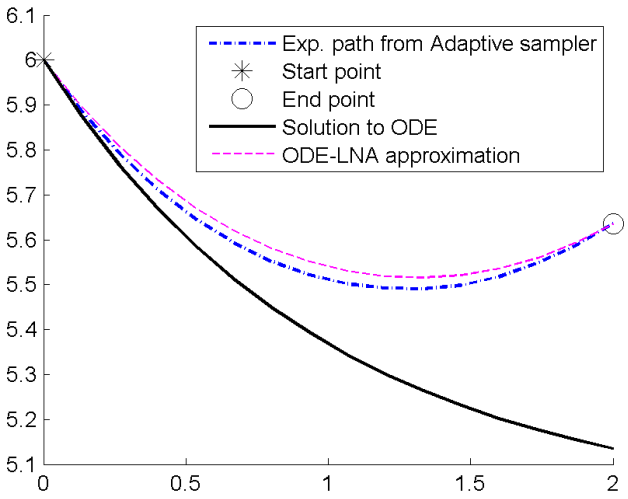


Figure 5.2: *The figure shows the expected path of the Cox-Ingersoll-Ross process in question in dash-dotted blue (Monte Carlo estimate obtained using the proposed, adaptive sampler), the solution to the corresponding ordinary differential equation in solid black, and the linear noise approximation of the expected path in dashed magenta.*

equation of course admitted attraction to one attractor alone, but the random part caused some trajectories to be caught in the field of attraction of the wrong attractor, and veer significantly from their proper course.

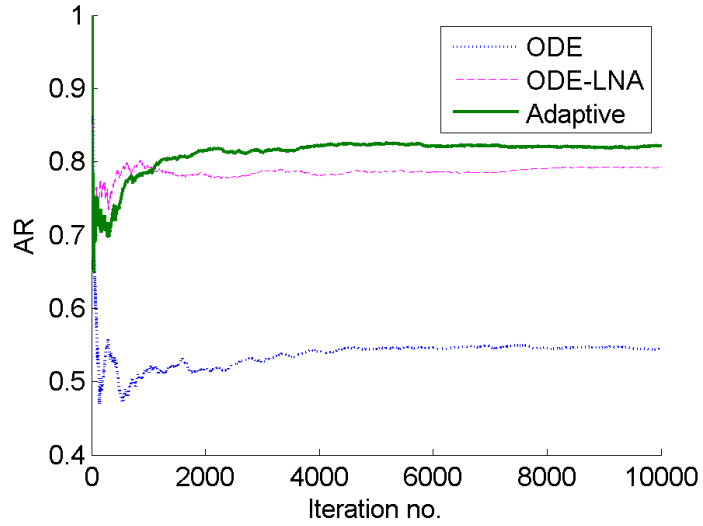
Initially, the sine approximation was used to approximate the expected path, but after a burn-in period of 3 000 iterations, the updating method is switched to recursively taking the mean of obtained realisations.

5.2.1 Easy case

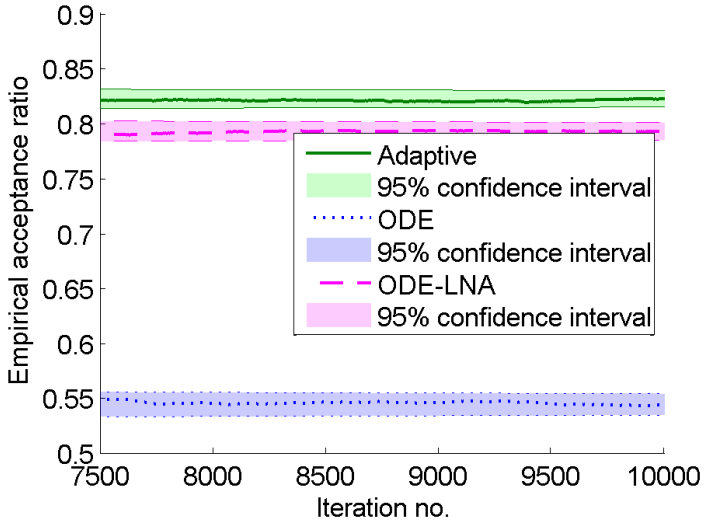
An easy, in some respect, path within the Lorenz model can be obtained by choosing a short time horizon of $T = 0.1$, and steering clear of bifurcations. Paths in this setting resemble the deterministic path. Such a path is shown in Fig. 5.4, where the deterministic path is shown in solid black, and the mean paths from the proposed sampler and the LNA sampler are shown in solid green and dashed magenta, respectively.

Fig. 5.5 shows an estimate of the evolution of the acceptance ratio for a single bridge to a specific end-point. It should be noted that the relatively few iterations of the simulations make for a large uncertainty of the acceptance ratio estimates. As such, the evolution of the acceptance ratio should be taken with a grain of salt for early iterations. The randomness of the simulated bridges, which are used to update the expected path, is so great that a number of 'poor' bridges in the beginning can lengthen the time to adequate convergence.

It can be readily seen in the aforementioned figures that the proposed sampler performs on a par with the LNA sampler; we see that the resulting expected paths are virtually indistinguishable and that the adaptive sampler performs no worse than the LNA sampler. Both the proposed sampler and the LNA sampler seem to be relatively unaffected by changes in the number of imputed points for the easy cases presented



(a) *The figure shows the evolution of the acceptance ratio for the entire run.*



(b) *Here, the evolution of the acceptance ratio over the last* $2\,500$ *iterations of the run is shown with the addition of the corresponding* $95\,\%$ *confidence intervals.*

Figure 5.3: *The figure shows the evolution of the acceptance ratio over time for the Cox-Ingersoll-Ross model. The estimates are obtained at each iteration as the number of accepted trajectories so far divided by the number of iterations. N.B. the great uncertainty of the estimates in the beginning, being based on very few iterations. Towards the end, however, one can clearly see a significant difference between the three samplers, with the proposed adaptive sampler performing best.*

here, and changes to 30 or 70 imputed points – instead of the 50 points shown here – yielded no discernible difference.

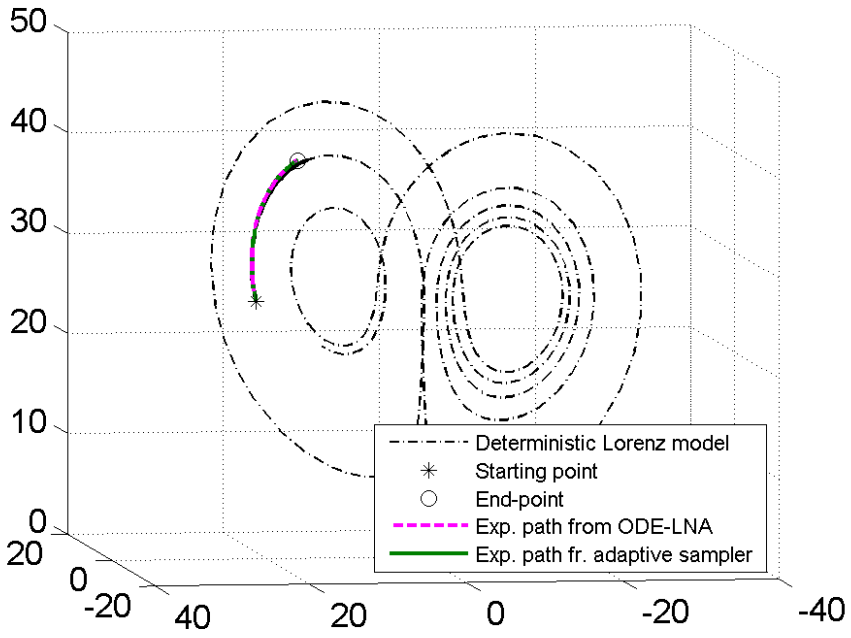


Figure 5.4: *The figure shows an example of a trajectory in the Lorenz model. The mean paths obtained from the proposed adaptive sampler and from the LNA sampler (Whitaker et al. 2016) are shown in solid green and dashed magenta, respectively, and they are virtually indistinguishable. The solution to the deterministic initial value problem is shown in solid black, and a long-time solution to the deterministic problem is shown as background in dash-dotted black. The starting point is marked by a star and the end-point by a circle.*

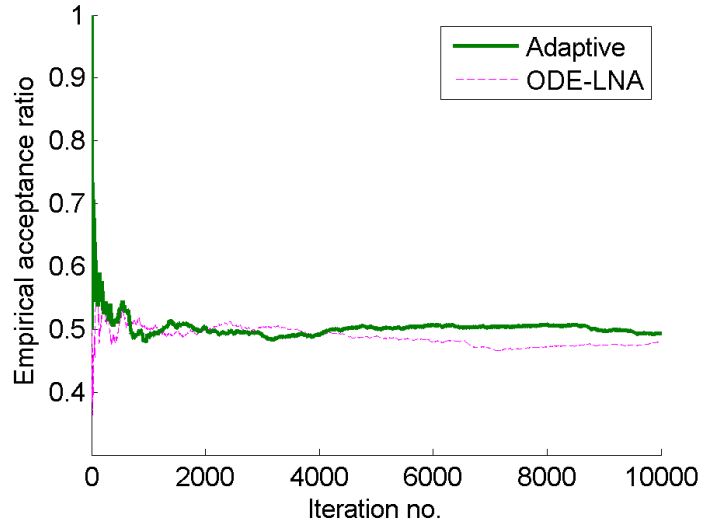
Since the proposed sampler uses the linear noise approximation as initial distribution, one would surmise that it would be more time consuming than the LNA sampler. Such an increase in time was negligible for the simulations herein presented.

5.2.2 Difficult case

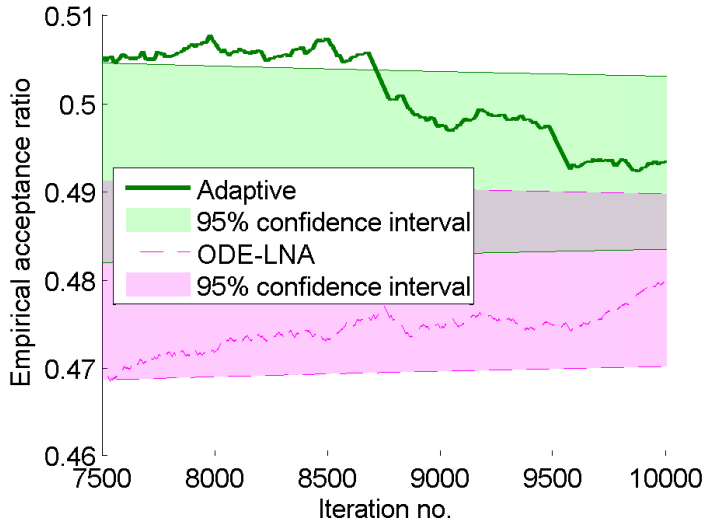
Placing the starting point close to the boundary between the fields of attraction of the two attractors, renders a more difficult test case, where the trajectories may be caught in the field of attraction of the wrong attractor and be swept far away from where the deterministic solution would take it. The time horizon was also increased to $T = 0.3$.

Fig. 5.6 shows the mean paths from both the proposed, adaptive sampler and from the LNA sampler for an end-point quite far from the deterministic solution. In this, admittedly rather extreme, case the LNA sampler yields quite poor results, whereas the adaptive sampler produces a much ($\sim 30\,\%$) higher acceptance ratio. Fig. 5.7b clearly shows this difference, and that the difference is, in fact, significant. The acceptance ratio of the adaptive sampler also seems to be rather constant throughout the chain, hovering at just over $40\,\%$ for all iterations after the initial erratic behaviour, where little can be said due to the uncertainty of the estimates. This is shown in Fig. 5.7.

Bridges were simulated from the starting point to 30 different end-points, with each



(a) *The figure shows the evolution of the acceptance ratio for the entire run.*



(b) *Here, the evolution of the acceptance ratio over the last* $2\,500$ *iterations of the run is shown with the addition of the corresponding* $95\,\%$ *confidence intervals.*

Figure 5.5: *The figure shows the evolution over time of the estimates of the acceptance ratio for one single example end-point for the easy case of the Lorenz model. The estimates over time are obtained for every iteration as the number of accepted draws up to the current iteration divided by the number of iterations. The estimated acceptance ratios are shown in solid green for the proposed, adaptive sampler and in dashed magenta for the LNA sampler (Whitaker et al. 2016). The simulations shown here were carried out with 50 imputed points, but no discernible difference can be found when decreasing the number to 30, or increasing the number to 70 imputed points. No significant increase in acceptance ratio can be observed when using the adaptive sampler.*

simulation consisting of $K = 10\,000$ iterations, and $m = 50$ data-points being imputed into the interval. Fig. 5.8 shows estimated acceptance ratios (i.e. the total number of accepted proposals divided by the total number of iterations) against the distance between the end-point and the end-point of the deterministic initial value problem. Here, it is readily seen that the proposed, adaptive sampler is far more robust than the LNA sampler, handling the errant end-points significantly better, whilst yielding much the same results close to the deterministic end-point.

The recursive formulations used to update the approximation of the expected path in the adaptive sampler are not particularly heavy, and time consumption does not seem to increase rampantly for errant end-points. The time required to perform the $10\,000$ iterations for the end-point shown in Fig. 5.6 with the adaptive sampler was but a fraction of a per cent higher than with the LNA sampler.

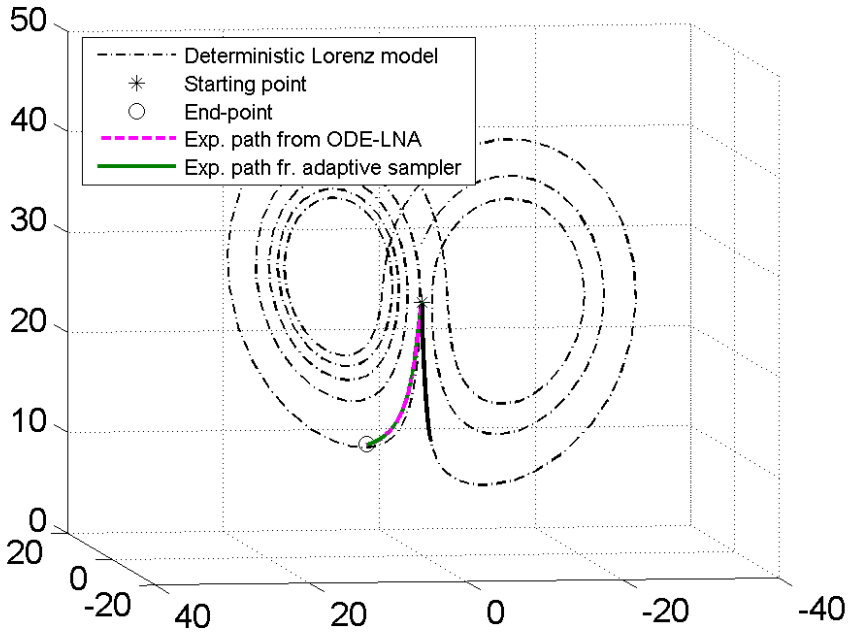
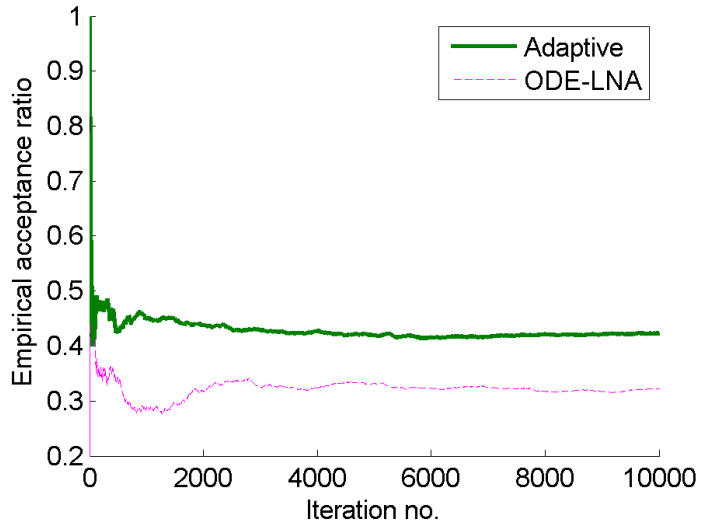
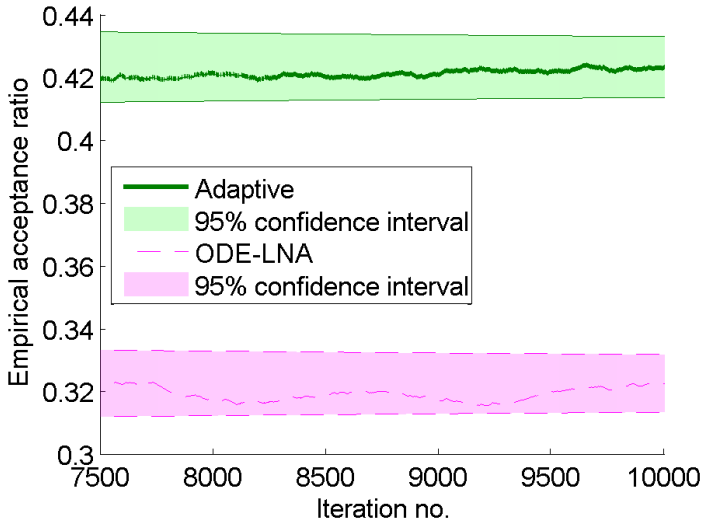


Figure 5.6: *The figure shows an example of a trajectory from the difficult case of the Lorenz model, veering from the course prescribed by the ordinary differential equation and into the field of attraction of the wrong attractor. The mean paths obtained from the proposed adaptive sampler and from the LNA sampler (Whitaker et al. 2016) are shown in solid green and dashed magenta, respectively, with no discernible difference. The solution to the deterministic initial value problem is shown in solid black, and a long-time solution to the deterministic problem is shown as background in dash-dotted black. The starting point is marked by a star and the end-point by a circle.*



(a) *The figure shows the evolution of the acceptance ratio for the entire run.*



(b) *Here, the evolution of the acceptance ratio over the last* $2\,500$ *iterations of the run is shown with the addition of the corresponding* $95\,\%$ *confidence intervals.*

Figure 5.7: *The figure shows the evolution over time of the estimated acceptance ratios for the difficult case of the Lorenz model shown in Fig. 5.6, where the end-point has gone far away from the ODE end-point and far into the field of attraction of the wrong attractor, i.e. a very difficult example. The estimates are obtained for every iteration as the number of accepted draws up to the current iteration divided by the number of iterations. The estimated acceptance ratios are shown in solid green for the proposed, adaptive sampler and in dashed magenta for the LNA sampler (Whitaker et al. 2016). In Fig. 5.7b, one can clearly see a significant difference between the proposed, adaptive sampler and the LNA sampler.*

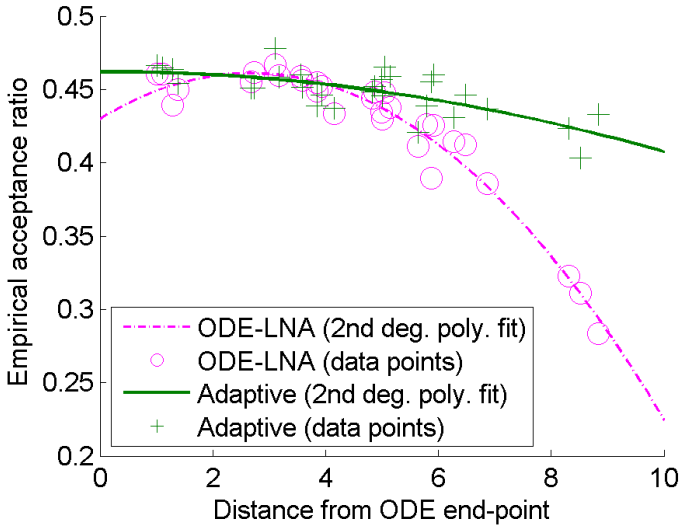


Figure 5.8: *The figure shows estimated acceptance ratios for the difficult set-up of the Lorenz model, based on $K = 10\,000$ iterations, for bridges for 30 different end-points against the distance of the end-points to the end-point of the deterministic problem. Ratios for the proposed, adaptive sampler is shown as green pluses, and ratios for the LNA sampler is shown as magenta-coloured circles. In order to illustrate the perceived trend of the data, second degree polynomials have been fitted to the obtained data – in a least-squares sense – and they are shown as a green line and a dash-dotted, magenta-coloured line for the adaptive and the ODE-LNA sampler, respectively. For end-points close to the deterministic end-point, the two samplers perform roughly equally, but as the distance to the ODE end-point increases, the LNA sampler deteriorates at a higher rate than the proposed, adaptive sampler.*

Chapter 6

Discussion

In this report, an algorithm for the simulation of bridge processes for stochastic differential equations has been presented. It is based on the concept of residual bridges from Whitaker et al. 2016, wherein a deterministic process is subtracted from the actual process, yielding a residual process, the dynamics of which one hopes to be linear enough for the application of a modified diffusion bridge. Whereas Whitaker et al. approximates the conditioned path by a deterministic system of ordinary differential equations, it is proposed herein that previous, accepted realisations be used to adaptively update the approximation of the conditioned, expected path, continually making it approach the true one, and thereby improve performance. The idea behind the proposed sampler is thus that the best proposal distribution has the correct expected path.

The efficiency of the proposed sampler depends greatly on the speed with which the approximation of the expected path converges to the true expected path. If only a few realisations are available to create the approximation, their effect on the path around which new proposals are created will be great, and the obtention of a series of 'bad', i.e. very erratic, realisations in the beginning of the run can seriously distort to the approximate expected path, negatively impacting the performance of the sampler. The rate of convergence, and the charting of the types of model for which convergence is quick enough, are areas that merit further research.

Hence, it is necessary to dull the effects of initial realisations, until there are enough for the law of large numbers to ensure a good approximation. The proposed sampler then models the remainder after the subtraction of the most up-to-date approximate path using regression with a suitable number of sine functions as basis for constructing a best approximation (the remainder is here seen as a part of a periodic function since it starts and ends at zero). It is important not to choose the number of sine functions too high. Naturally, one must not choose more basis functions than data points, but even a number close to the number of imputed points seemed to retain too many features of the noisy realisations to be of use in the desired, dulling manner. It is natural to surmise that the conditional expected path will bear some resemblance to the unconditional expected path, and that it is possible to look upon this for guidance as to how many basis functions are necessary.

A recursive regression is then easy to implement. This approximation, however, introduces bias, even in the limit, so after a suitable number of iterations, i.e. enough iterations so that the erratic behaviour of individual realisations has little impact upon the approximation, it is prudent to switch the updating method to a regular (recursively formulated) mean to ensure proper convergence. One may, for example, look at how

much each iteration affects the approximation and switch to taking the mean when the changes are small enough. For the Cox-Ingersoll-Ross and Lorenz models of the simulation study, the switch was carried out after $2\,000$ and $3\,000$ iterations, respectively.

The sampler was tested on two models: the Cox-Ingersoll-Ross (CIR) model and the stochastic Lorenz model, both capable of exhibiting significant non-linearities. For the CIR model, the proposed sampler performed well compared to the constructs in Whitaker et al. 2016, appearing more robust, both in terms of number of imputed values and in terms of how far from the deterministic end-point the end-point of the bridges were. These two considerations may be seen as two sides of the same coin. The actual distance from the deterministic path seemed less important than 'how non-linear' the expected path were. After all, less is required of an approximation between closely spaced points, so the linear noise approximation should yield better results for this case.

The stochastic Lorenz model is a three-dimensional, chaotic model that can be quite difficult to handle. The sampler were tested on two different cases: one fairly simple, and one more difficult, near a bifurcation. Also for this model, the proposed sampler managed to perform as well as, or better than, the samplers from Whitaker et al. 2016. For a substantial veer from the deterministic path, e.g. a trajectory entering the field in which the 'wrong' attractor dominates, the non-linearity of the expected path becomes very strong, and the LNA sampler performs poorly, as expected. The proposed sampler, however, weathers this problem without much difficulty, which is to be expected as the proposed sampler should eventually attain the true expected path, regardless of how non-linear it is.

It is reasonable to assume that the proposed sampler should be more computationally intensive than its competitors – after all, since it uses the LNA as an initial distribution, it does everything that they do and more. However, it is known that the recursive formulations of both the updating of the mean and the updating of the regression are light, only having a cost of $O(N)$, whereby one is led to believe that computational cost should not be too great a problem. Quick examinations of the time consumption of the simulations in Chapter 5, which, one must concede, could have been carried out more rigorously, showed but a marginal increase. This, however, is an area wherein more work remains.

There are other simulation methods, see e.g. Guided proposals from Schauer, Meulen and Zanten 2013, based on quite different methods than the residual bridges considered herein, which have not been considered. Another area that merits further investigation is the parameter estimation problems from too sparse data, mentioned in the introduction. An interesting query would be how this sampler can be used for that purpose. This has not been treated in this project. A previously examined method has used a Gibbs sampling framework, alternating between sampling bridges (applying modified diffusion bridges to all intermediate intervals) and sampling parameters, see Jensen et al. 2012, and it would be interesting to see whether the proposed sampler could be efficient in such a context.

In conclusion, the proposed method is on a par with earlier methods, presented in Whitaker et al. 2016, for easy cases, and significantly improves performance for difficult cases. The proposed method is thus more robust, whilst only increasing computing times marginally.

Bibliography

Andrieu, Christophe and Johannes Thoms (2008). 'A tutorial on adaptive MCMC'. In: *Statistics and computing* 18.4, pp. 343–373.

Atchadé, Yves et al. (2011). 'Adaptive Markov chain Monte Carlo: theory and methods'. In: *Bayesian Time Series Models*. Ed. by David Barber, A. Taylan Cemgil and SilviaEditors Chiappa. Cambridge University Press, pp. 32–51. DOI: 10.1017/CBO9780511984679.003.

Bai, Yan, Gareth O Roberts and Jeffrey Seth Rosenthal (2009). *On the containment condition for adaptive Markov chain Monte Carlo algorithms*. Technical report: *University of Warwick. Centre for Research in Statistical Methodology*. URL: http://probability.ca/jeff/ftpdir/yannew.pdf.

Bengtsson, Thomas, Chris Snyder and Doug Nychka (2003). 'Toward a nonlinear ensemble filter for high-dimensional systems'. In: *Journal of Geophysical Research: Atmospheres* 108.D24.

Björk, Tomas (2009). *Arbitrage theory in continuous time*. Oxford university press.

Carleson, Lennart (1966). 'On convergence and growth of partial sums of Fourier series'. In: *Acta Mathematica* 116.1, pp. 135–157.

Cox, John C, Jonathan E Ingersoll Jr and Stephen A Ross (1985). 'A theory of the term structure of interest rates'. In: *Econometrica: Journal of the Econometric Society*, pp. 385–407.

Debnath, Lokenath and Piotr Mikusinski (1999). *Introduction to Hilbert Spaces with Application, 1999*.

Delyon, Bernard and Ying Hu (2006). 'Simulation of conditioned diffusion and application to parameter estimation'. In: *Stochastic Processes and their Applications* 116.11, pp. 1660–1675.

Durham, Garland B and A. Ronald Gallant (2002). 'Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes'. In: *Journal of Business & Economic Statistics* 20.3, pp. 297–338. URL: http://dx.doi.org/10.1198/073500102288618397.

Fearnhead, Paul, Vasilieos Giagos and Chris Sherlock (2014). 'Inference for reaction networks using the linear noise approximation'. In: *Biometrics* 70.2, pp. 457–466.

Feller, William (1951). 'Two singular diffusion problems'. In: *Annals of mathematics*, pp. 173–182.

Givens, Geof H and Jennifer A Hoeting (2012). *Computational statistics*. Wiley Series in Computational Statistics. John Wiley & Sons.

Hastings, W Keith (1970). 'Monte Carlo sampling methods using Markov chains and their applications'. In: *Biometrika* 57.1, pp. 97–109.

Jensen, Anders Chr et al. (2012). 'Markov chain Monte Carlo approach to parameter estimation in the FitzHugh-Nagumo model'. In: *Physical Review E* 86.4, p. 041114.

Karatzas, Ioannis and Steven Shreve (1998). *Brownian Motion and Stochastic Calculus*. 2nd. Vol. 113. Graduate Texts in Mathematics. Springer Science & Business Media.

Keller, Hannes (1996). *Attractors And Bifurcations Of The Stochastic Lorenz System*. Technical Report 389. Institut für Dynamische Systeme, Universität Bremen.

Kloeden, Peter E. and Eckhard Platen (1992). *Numerical solution of stochastic differential equations*. Vol. 23. Stochastic Modelling and Applied Probability. Springer-Verlag Berlin Heidelberg.

Langevin, Paul (1908). 'Sur la théorie du mouvement brownien'. In: *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 146, pp. 530–533.

Lindström, E., H. Madsen and J.N. Nielsen (2015). *Statistics for Finance*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. ISBN: 9781482229004. URL: https://books.google.se/books?id=xXR3CAAAQBAJ.

Lindström, Erik (2007). 'Estimating parameters in diffusion processes using an approximate maximum likelihood approach'. In: *Annals of Operations Research* 151.1, pp. 269–288. ISSN: 1572-9338. DOI: 10.1007/s10479-006-0126-4. URL: http://dx.doi.org/10.1007/s10479-006-0126-4.

— (2012). 'A regularized bridge sampler for sparsely sampled diffusions'. In: *Statistics and Computing* 22.2, pp. 615–623. ISSN: 1573-1375. DOI: 10.1007/s11222-011-9255-y. URL: http://dx.doi.org/10.1007/s11222-011-9255-y.

Lo, Andrew W. (1988). 'Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data'. In: *Econometric Theory* 4.2, pp. 231–247. ISSN: 02664666, 14694360. URL: http://www.jstor.org/stable/3532294.

Lorenz, Edward N (1963). 'Deterministic nonperiodic flow'. In: *Journal of the atmospheric sciences* 20.2, pp. 130–141.

Lyons, Simon MJ (2013). 'Introduction to stochastic differential equations'. In: URL: http://homepages.inf.ed.ac.uk/s0978702/introsde.pdf.

Metropolis, Nicholas et al. (1953). 'Equation of state calculations by fast computing machines'. In: *The journal of chemical physics* 21.6, pp. 1087–1092.

Øksendal, B. (1998). *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg. ISBN: 9783662036204. URL: https://books.google.se/books?id=gizqCAAAQBAJ.

Papaspiliopoulos, Omiros and Gareth Roberts (2012). 'Importance sampling techniques for estimation of diffusion models'. In: *Statistical methods for stochastic differential equations* 124, pp. 311–340.

Papaspiliopoulos, Omiros, Gareth O Roberts and Osnat Stramer (2013). 'Data augmentation for diffusions'. In: *Journal of Computational and Graphical Statistics* 22.3, pp. 665–688.

Pedersen, Asger Roer (1995). 'A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations'. In: *Scandinavian Journal of Statistics* 22.1, pp. 55–71. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4616340.

Schauer, Moritz, Frank van der Meulen and Harry van Zanten (2013). *Guided proposals for simulating multi-dimensional diffusion bridges*. To appear in: *Bernoulli, 2017*.

Shiryayev, A.N. and R.P. Boas (1984). *Probability*. Graduate texts in mathematics. Springer. ISBN: 9783540908982. URL: https://books.google.se/books?id=SiWwQgAACAAJ.

Whitaker, Gavin A. et al. (2016). 'Improved bridge constructs for stochastic differential equations'. In: *Statistics and Computing*, pp. 1–16. ISSN: 1573-1375. DOI: 10.1007/s11222-016-9660-3. URL: http://dx.doi.org/10.1007/s11222-016-9660-3.