

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

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/164408>

Copyright and reuse:

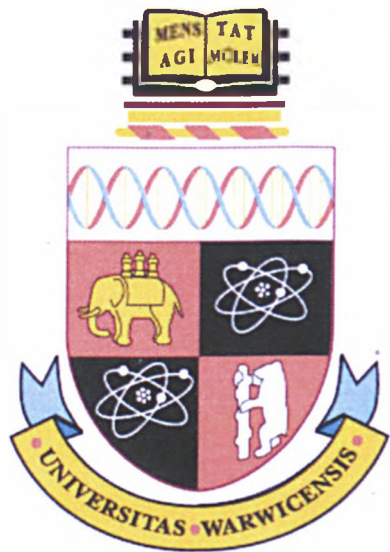
This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

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



**Fitting Stochastic Differential Equations to Molecular
Dynamics Data**

by

Yvo Pokern

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Institute of Mathematics

February 2007

AUTHOR: **Yvo Pokern** DEGREE: **Ph.D.**

TITLE: **Fitting Stochastic Differential Equations to Molecular Dynamics Data**

DATE OF DEPOSIT: ... 31.8.2006

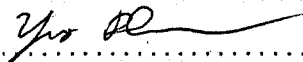
I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I agree that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries, subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

"Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: 

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

.....
.....
.....
.....

Contents

List of Figures	vi
Acknowledgments	ix
Declarations	x
Abstract	xi
Chapter 1 Introduction	1
1.1 Molecular Dynamics	1
1.2 Fitting Stochastic Differential Equations	2
1.3 Failure of Simple Estimators	4
1.3.1 Example I: Harmonic Oscillator	4
1.3.2 Example II: stochastic growth	5
1.3.3 Maximum likelihood estimation for the harmonic oscillator	5
1.4 Statisticians' and Practitioners' Approaches	6
1.5 Challenges and Cross-links – Conclusion	7
Chapter 2 Basics of Molecular Dynamics	8
2.1 Introduction	8
2.2 MD Force Field and Code	8
2.2.1 Protein model and force field	8
2.2.2 Implementation and Parameterisation	9
2.2.3 Verification of Force expressions	12

2.2.4	The Integrator	14
2.2.5	Lennard-Jonesium	19
2.2.6	Water	19
2.2.7	Metastability	21
2.3	Butane	22
2.4	Conclusion	22
Chapter 3 First Order Stochastic Differential Equations		26
3.1	Introduction	26
3.2	Continuous Time Path fitting	27
3.2.1	Estimating Diffusivity	27
3.2.2	Estimating Drift Parameters	29
3.3	Numerical Implementation	30
3.3.1	Diffusion Coefficient	31
3.3.2	Numerical Validation: Pitfalls	32
3.3.3	Drift Parameters: Polynomial Potential	37
3.3.4	Drift Parameters: Trigonometric Potentials	38
3.3.5	Decay of Variance	44
3.4	Application to Butane data	45
3.5	The Two scale Potential	47
3.6	Conclusions	48
Chapter 4 Second Order Stochastic Differential Equations		49
4.1	Overview	49
4.2	Introduction	50
4.2.1	Literature review	54
4.3	Model Problems	56
4.3.1	Model Problem I: Stochastic Growth	57
4.3.2	Model Problem II: Harmonic Oscillator	57
4.3.3	Model Problem III: Oscillator with trigonometric potential	58
4.4	Euler Statistical Model	58

4.4.1	Statistical Model	58
4.4.2	Model Problem I	59
4.4.3	Analysis of why the missing data method fails	60
4.5	Improved statistical model	63
4.5.1	Path Sampling	64
4.5.2	Estimating diffusion coefficient and missing path	67
4.6	Drift Estimation	71
4.6.1	Overview	71
4.6.2	Drift parameters from \mathcal{L}_E	71
4.6.3	Drift parameters from \mathcal{L}_{IT}	73
4.6.4	Numerical Check: Drift	73
4.6.5	Why the Model fails for the drift parameters	75
4.6.6	Analysis of Drift Estimation Failure	75
4.6.7	Conclusion for Drift Estimation	78
4.7	The Gibbs Loop	79
4.7.1	Overview	79
4.7.2	The Algorithm	80
4.7.3	Combining MLE and Langevin estimators in a Gibbs Sampler	84
4.8	Application to Molecular Conformational Dynamics	90
4.8.1	Molecular Dynamics	90
4.8.2	Fitting	92
4.8.3	Limitations	93
4.9	Conclusions	95
Chapter 5 Nonparametric Estimation for Diffusion Processes		97
5.1	Overview	97
5.2	Introduction	97
5.3	The Gradient Case	101
5.3.1	Statisticians' Approach	101
5.3.2	Practitioners' Approach	102
5.3.3	Specialising to 1D - finite T results	103

5.3.4	Extension to higher dimensions?	104
5.4	Drift Estimation for Reversible Processes	104
5.5	Drift Estimation for Second Order Langevin Equations	106
5.5.1	Direct Variational Approach	106
5.5.2	Langevin in the general Framework	108
5.5.3	Nonparametric Estimation of $b(x)$	109
5.5.4	Nonparametric Estimation of $(V(q), \beta)$	110
5.6	Estimating the Diffusion Coefficient	112
5.6.1	Statisticians' Approach	112
5.6.2	Autocorrelation Function	112
5.7	Relationship to existing literature	113
5.8	Numerical Experiments	115
5.8.1	Introduction	115
5.8.2	Empirical and MLE-induced Probability Densities	116
5.8.3	Parametric Estimation	118
5.8.4	Correlation of estimated drift parameters	124
5.8.5	Comparing autocorrelations	128
5.8.6	Misspecified model – multiplicative noise	132
5.9	Conclusions and Future Work	134

List of Figures

2.1	Atom Bond	10
2.2	Bond Angle	10
2.3	Dihedral Angle	11
2.4	Van der Waals Interaction	11
2.5	Electrostatic Coulomb Interaction	12
2.6	Force errors: BS, BA, DA	13
2.7	Force errors: vdW, Coulomb	14
2.8	Comparison of Integrators	15
2.9	Verlet Integrator Verification	16
2.10	Semifolded intermediary configuration	17
2.11	Verlet Integrator Verification 2	18
2.12	RMS Energy deviations	18
2.13	Total Energy fluctuations	19
2.14	Radial Distribution function for Argon	20
2.15	Butane Dihedral Angle Trajectory	23
2.16	Langevin Dynamics for Butane – Invariant Measure and Empirical Density	24
3.1	σ Sampling Error	34
3.2	σ Sampling Error - logarithmic	34
3.3	Variance Error for LNC RNG	36
3.4	Mersenne Twister – Variance	36
3.5	θ_5 fit for polynomial potential	38

3.6	θ_3 fit for polynomial potential	39
3.7	Trigonometric Potential and Sample Path	40
3.8	Convergence of $\hat{\theta}_1$	40
3.9	Trigonometric potential – unfavourable case	41
3.10	Low $\hat{\theta}_1$ Errors in unfavourable case	42
3.11	Predicted 1st Order Errors – favourable case	43
3.12	Predicted 1st Order Errors – unfavourable case	44
3.13	Fitting Trigonometric Potentials to Butane – Coefficients	46
3.14	Different Potentials at Different k	47
4.1	Estimates of σ using Euler Model for Model Problem I. Top row: fully observed process; bottom row: partially observed process.	61
4.2	Estimates of σ using the \mathcal{L}_{IT} Model for Model Problem I. Top row: fully observed process; bottom row: partially observed process.	69
4.3	Estimates of σ using the \mathcal{L}_{IT} Model for Model Problem II. Top row: fully observed process; bottom row: partially observed process.	70
4.4	Drift estimation for Model Problem II, using \mathcal{L}_{IT}	74
4.5	Typical sample path for Model Problem III, $T = 500$	81
4.6	Whole loop estimation for Model Problem III: $T = 500$	82
4.7	Whole loop estimation for Model Problem III: $T = 50$	83
4.8	Whole loop estimation for Model Problem III: $T = 500$	85
4.9	Comparing Hybrid and all-sampling Algorithms	89
4.11	MD Samplepath: Butane	91
4.10	Sketch of Dihedral Angle	91
4.12	Convergence for fitted MD path with subsampling	93
4.13	PDFs resulting from fitted potentials for different orders of trigonometric potential - Shaded regions display posterior variance	95
5.1	probability density functions from one particular samplepath	117
5.2	Convergence as $\Delta t \rightarrow 0$ of drift parameters for MLE	120
5.3	Convergence as $\Delta t \rightarrow 0$ of diffusion parameter for MLE	121

5.4	Convergence as $\Delta t \rightarrow 0$ of drift parameters for MDE	122
5.5	Convergence as $\Delta t \rightarrow 0$ of diffusion parameter for MDE	123
5.6	Confluence of MD estimates for perfect histogram data	123
5.7	Convergence as $\Delta t \rightarrow 0$ of drift parameters for Practitioners' estimator	124
5.8	Deviations of drift Parameter θ_3 from mean, $T_f = 20480$	125
5.9	Correlations of drift parameter deviations	126
5.10	Correlations of drift parameter deviations for \tilde{A} vs. MLE	127
5.11	Comparing correlations of estimated drift parameters for MDE, Practitioners' and MLE estimated drift parameters	128
5.12	Comparing autocorrelations for MDE and MLE estimated parameters .	129
5.13	Comparing autocorrelations for MDE and MLE estimated parameters, true σ for MDE	130
5.14	Autocorrelations for damped-driven harmonic oscillator	131
5.15	Comparing induced PDFs for MLE and MDE	133
5.16	Comparing induced autocorrelations for MLE and MDE	134

Acknowledgments

First and foremost, I would like to thank my supervisor, Andrew Stuart, for his generous support, both material and immaterial, and unfailing encouragement. Without his continual motivation and knack for finding fruitful research problems, this work would not have been produced.

Thanks are also due to Mark Rodgers from the Warwick Chemistry department, Jochen Voss and Petter Wiberg from Warwick Mathematics who participated in valuable and stimulating discussions. Fellow students Richard Fielding, Jim MacDonald, Oliver Tearne, Paul Wheeler, David White and Kostas Zygalkakis have helped to broaden my mathematical background and satisfied my desire to poke my nose into other people's research problems.

Finally, the Anchor House community, in particular Sharae and Milena deserve thanks for helping to maintain my cultural and social life while immersing myself in science.

Declarations

The work in chapters 1, 2 and 3 has been produced solely by the author under the supervision of his thesis supervisor.

Chapter 4 is based on a joint paper, [61], with Andrew Stuart and Petter Wiberg. All the numerics in that chapter have been implemented, tested and put into application by the author of the current thesis. Initial observations on incorrect estimates for the quadratic variation as well as on a non-systematic way of arriving at \mathcal{L}_{IT} were made by Stuart and Wiberg who also provided a large part of the literature review. The analytic understanding of drift estimation and the compound algorithms applied in this context were contributed by the author of this thesis.

Chapter 5 is based on a joint paper in preparation with Andrew Stuart and Eric Vanden-Eijnden, [15], in which all the numerics are the thesis' author's contribution. While the replacement of the stochastic integral using the Kolmogorov equation and the attendant consideration of reversible processes is due to Stuart and Vanden-Eijnden, the extension to the second order Langevin case as well as results for finite final time have been contributed by the author of the thesis.

The material in this thesis is submitted for a degree to the University of Warwick only and has not been submitted to another university.

Abstract

The thesis consists of three main parts. Firstly, a molecular dynamics and potential energy minimisation package that has been implemented is described in detail. All potential and force interactions are described and tested successfully. Compound tests on minimal energies for clusters of water molecules, the radial distribution function for liquid argon and the equilibrium distribution for the dihedral angle in Butane under Langevin dynamics are performed and the presence of multiple time scales is noted for Butane as well as for a simplified protein model due to Grubmüller and Tavan.

Secondly, fitting stochastic differential equations (SDEs) to time series is studied. Initially, I consider the well-understood case of non-degenerate diffusions, where all components of the process are driven directly by Brownian motion. An SDE with constant diffusivity and trigonometric force expression is fitted to trajectories obtained from simulations of Butane by maximum likelihood methods and fitted diffusion and drift parameters depend strongly on the timescale considered. Hypoelliptic diffusion processes are considered next. Here, the unexpected failure of simple estimators necessitates the use of carefully chosen approximate likelihoods. For the case of only partial observations being available, a compound algorithm is designed and numerically seen to be asymptotically consistent. It is applied to the same Butane sample path and found to equilibrate, although the fitted SDE fails to reproduce the free energy landscape.

Thirdly, connections between maximum likelihood estimators (MLEs) and practitioners' methods are investigated. Analytical links are found for reversible processes and for second order Langevin processes. In the case of 1D processes, MLE and practitioners' methods for the drift are found to yield estimators identical up to lower order terms even for finite times of observation.

Chapter 1

Introduction

In order to give an overview of what will be treated in this thesis, this chapter will briefly touch on current challenges in molecular dynamics simulations and the motivation for fitting stochastic differential equations to trajectories from molecular dynamics simulations in section 2 and 3. Section 3 then goes on to describe basic techniques of parameter estimation and section 4 describes challenges arising from only partial observations being available. Finally, section 5 will comment on relations between practitioners' and statisticians' methods to estimate drift and diffusion parameters.

1.1 Molecular Dynamics

Molecular Dynamics is currently faced with a computational bottleneck: There is a gap of several orders of magnitude between the total time a direct, atomically resolved simulation of a macromolecule of biological interest can cover and the timescales at which biologically interesting dynamics occur. Some computational biologists emphasise that minimal energy conformations of proteins are a good test for forcefields and are sceptical biologically meaningful direct molecular dynamics simulations are at all feasible given today's computational resources. Nonetheless, packages for direct molecular dynamics simulation have been developed for a long time, with [17], [20] and [22] being the original publications for CHARMM, AMBER and ECEPP respectively, all of which were written in the 1970s and 1980s.

While chemists tend to stress the importance of the forcefield and the molecules used to train the forcefield, recent interest in the applied mathematics community has been focused on dynamics, in particular on how to extract or compute effective dynamics. One point which is thought to be characteristic of proteins is the existence of metastable states, states which are separated by a large energy barrier which is seldom crossed at room temperature. Reproducing this qualitative feature with a very simple protein model was the main focus of [32] and in fact [31].

1.2 Fitting Stochastic Differential Equations

It is not only the generation of molecular dynamics data which currently presents a computational bottleneck, other difficulties arise when attempting to interpret this data. The extraction of physically meaningful essential dynamics has been a focus for some time and one way of extracting this information that has been suggested is to fit stochastic processes, or, given the continuous time nature of these processes, stochastic differential equations to time series from molecular dynamics. If the fitted SDE is well-chosen, the parameters represent physically meaningful quantities and can thus be viewed as the extracted dynamical information. Early work in this direction can be found in [32], [59]. A more elaborate approach is taken by Hummer in [34] using multiplicative noise, and by Schütte and coworkers in [35] using a hidden Markov model to switch between different SDEs.

One might ask why one should attempt to fit a stochastic process to an entirely deterministic Hamiltonian system. Practitioners generally quote the large number of particles involved and vaguely appeal to concepts of statistical physics rather than mathematically rigorous ergodic theory. In fact, in the context of a distinguished particle in a heat bath it can be shown that in the limit of large particle numbers, the distinguished particle's trajectory converges to those of a certain SDE in a rather weak sense ([27], [63]). Following a different line of argument, one can appeal to the different timescales involved: fast, oscillatory movement at the atomic level and relatively slow movement at the conformational level. This scale separation can also be used to

rigorously justify stochastic behaviour if the fast driving process is chaotic, see [25]. An overview of extracting effective dynamics from a mathematical perspective can be found in [10].

Following this motivation, the current thesis will be concerned with fitting stochastic differential equations to trajectory data from molecular dynamics simulations. The first step towards this, however, has to be the fitting to trajectories that are actually generated from an SDE of the type to be fitted. Only if this fitting is successful, one can think of applying the algorithms to molecular dynamics data.

To conclude this section, I briefly review the maximum likelihood estimator technique for fitting parameters in a stochastic differential equation. Consider the equation

$$dx = \Theta A(x)dt + \Sigma dW \quad (1.1)$$

where $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ is the solution of the SDE, $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are suitably well-behaved force functions, $\Theta \in \mathbb{R}^{n \times m}$ are parameters and Σ is the invertible diffusivity matrix. The likelihood is then given via the Girsanov formula relating the probability density \mathbb{P}_Θ on path space for the process specified by (1.1) to that of Wiener measure \mathbb{P} with diffusivity Σ . This is done by specifying the Radon-Nikodym derivative which is given as

$$\frac{d\mathbb{P}_\Theta}{d\mathbb{P}} = \exp \left(\int_0^T \Sigma^{-1} \Theta A(x(s)) \Sigma^{-1} dx_s - \frac{1}{2} \int_0^T \|\Sigma^{-1} \Theta A(x(s))\|^2 ds \right)$$

Given a finite piece of trajectory, $\{x(t)\}_{t \in [0, T]}$, one can maximise the likelihood of the given path using the drift coefficients Θ provided *independently of* Σ by the estimator

$$\hat{\Theta} = \left(\int dx_s \otimes A(x(s)) \right) \left(\int A(x(s)) \otimes A(x(s)) ds \right)^{-1}. \quad (1.2)$$

if the diffusion process is suitably ergodic. For non-invertible Σ some results are known if the process is hypoelliptic. In the case of linear force functions it has been shown that (1.2) is still viable and there are numerical indications presented in this thesis that this extends to suitable force functions $A(\cdot)$. If only some of the entries of Θ are to be estimated, however, knowledge of Σ can enter the estimator for Θ and can lead to ill-conditioning as highlighted in the next section, see in particular 1.3.3.

1.3 Failure of Simple Estimators

This introduction will briefly highlight some examples studied in more depth in the subsequent chapters. Where full details are not given in the introduction, they will of course be provided in subsequent chapters.

1.3.1 Example I: Harmonic Oscillator

Following the example of the distinguished particle in the heat bath, let's consider a very simple model: the harmonic oscillator with white noise forcing.

$$\ddot{x} + \gamma\dot{x} + Dx = \sigma\dot{W}$$

This second order SDE should be interpreted in the following sense:

$$\begin{aligned} dq &= p dt \\ dp &= -Dq dt - \gamma p dt + \sigma dW \end{aligned} \tag{1.3}$$

where W is standard Brownian motion and all quantities are scalar.

The initial problem is to estimate the parameters D , γ and σ given a finite number of observations $q_i, p_i, i \in \{1, \dots, N\}$ at equidistant times t_i .

Several observations lead to consider the problem of only partially available data, though:

On the side of the fitted model, it is clear from (1.3) that the spatial component, q will be $C^{1,\alpha}$ for any $\alpha \in [0, \frac{1}{2})$ whereas the momentum component, p , will be rougher.

On the other hand, taking data from a molecular dynamics simulation where a molecule may be modelled as a Hamiltonian dynamical system, it is clear that this data will be smooth, provided the potentials used are sufficiently well-behaved. Therefore, as the spacing between the observation times $\Delta t = t_i - t_{i-1}$, goes to zero, and the final observation time goes to infinity, convergence can only be expected in some weak sense. Numerically, this manifests itself in experiments with parameter estimators for σ which are based on quadratic variation and suffer from $\sigma \rightarrow 0$ as $\Delta t \rightarrow 0$.

Finally, for molecular dynamics simulations, velocity data is not always available at the required times (typically, a Stoermer-Verlet scheme is used which delivers $p_{n+\frac{1}{2}}$

rather than p_n). If simple interpolation formulae are used this is tantamount to numerical differentiation which can lead to strange behaviour of estimators for σ as next subsection shows.

1.3.2 Example II: stochastic growth

Consider model problem I from chapter 4:

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= \sigma \dot{W} \end{aligned} \quad (1.4)$$

Using a straightforward numerical differentiation formula to estimate the unobserved velocities

$$\hat{p}_n = \frac{q_{n+1} - q_n}{\Delta t} \quad (1.5)$$

which corresponds to a maximum likelihood estimator arising from an explicit Euler statistical model, the estimate for the diffusion coefficient σ is biased. In fact, it is shown in subsection 4.4.3:

$$\hat{\sigma}^2 \rightarrow \frac{2}{3}\sigma^2 \text{ as } T \rightarrow \infty. \quad (1.6)$$

Thus, numerical differentiation can lead to completely wrong estimates. For the harmonic oscillator (1.3), though, even worse is true:

1.3.3 Maximum likelihood estimation for the harmonic oscillator

Suppose observations $q_i, p_i, i \in \{1, \dots, N\}$ of the harmonic oscillator (1.3) are available at equidistant times $t_i = i\Delta t$ and we wish to estimate γ, D, σ from these observations using a maximum likelihood estimator. For the straightforward maximum likelihood estimator it is then possible to show that the drift parameters are similarly off-track. In fact,

$$\mathbb{E}\hat{D} = \frac{1}{4}D \quad (1.7)$$

$$\mathbb{E}\hat{\gamma} = \frac{1}{4}\gamma \quad (1.8)$$

holds!

It is the hypoelliptic nature of these problems which forces a certain structure on the estimators, taking into account the propagation of noise into the smooth components. This propagation results in ill-conditioned statistical models which necessitate careful selection of drift estimators if the compound algorithm is to be asymptotically consistent. A full exposition of these issues can be found in chapter 4.

1.4 Statisticians' and Practitioners' Approaches

Naturally, there is statistical literature about fitting stochastic differential equations to time series data, e.g. [50] and more recently [62] provide overviews. The statistical literature frequently assumes that estimating diffusion coefficients is easy (e.g. Kutoyants, [62] completely excludes the problem from consideration) arguing that it can be estimated from an arbitrary short piece of continuous-time trajectory using quadratic variation.

In the present application, however, this argument is not satisfactory because the processes only approximately behave like diffusion processes and their behaviour changes on the very shortest timescales.

From a physicist's point of view, estimating drift parameters is easy, provided one is given a sufficiently long piece of trajectory, assuming the system is in thermodynamic equilibrium in the canonical measure, as one can simply use the invariant measure to infer the drift coefficients, provided the temperature of the system is known. Estimating the diffusion coefficient is rather more difficult and there does not seem to be a canonical way of doing this as Hummer ([34]) points out.

It is thus of interest to link statisticians' and physicists' approaches to estimating parameters and some links between, firstly, maximum likelihood estimators and the fitting of the invariant measure, and, secondly, quadratic variation and fitting of the Laplace transform of the spatial autocorrelation are indeed found and described in Chapter 5.

1.5 Challenges and Cross-links – Conclusion

Having highlighted the continuing challenge of biologically meaningful molecular dynamics simulation and the fitting of stochastic differential equations as a means of extracting effective dynamics, some problems posed by those fitting procedures have been highlighted. Also, the cross-links of methods used traditionally in statistics and those used by physicists and chemists have been touched upon, thus summarising the main issues to be dealt with in this thesis.

Chapter 2

Basics of Molecular Dynamics

2.1 Introduction

In order to have a source of molecular dynamics data which is completely transparent, a simple MD code has been developed, implementing the absolutely essential features of CHARMM, [17].

The code has slowly grown into a multi-threaded C++ molecular dynamics and potential minimisation code with simple 3D visualisation routines (using GLUT). It uses essentials of the CHARMM force field as well as offering some alternative ad-hoc forcefields. The second section of this chapter will describe the forcefield and its implementation as well as various tests to which the code has been subjected. Integrators and energy conservation as well as compound tests such as properties of Lennard-Jonesium, minimal energies of water clusters and a simple protein model will also be studied in that section. The third section will focus on Butane and the observed metastability in this simplification of the Grubmüller/Tavan model.

2.2 MD Force Field and Code

2.2.1 Protein model and force field

The basics of molecular dynamics modelling can be found in [47] or the more recent book by Schlick, [54]. A Protein is simplistically modelled here as a system of N mass

points with mass m_i , whose position $q_i \in \mathbb{R}^3$ as a function of time is governed by the Hamilton equations of motion:

$$\begin{aligned} \dot{q}_{i,j} &= \frac{\partial H}{\partial p_{i,j}} \quad i \in \{0, \dots, N\} \\ \dot{p}_{i,j} &= -\frac{\partial H}{\partial q_{i,j}} \quad j \in \{1, 2, 3\} \end{aligned}$$

where $q_{i,j}$ refers to the j th coordinate of the i th atom. Of course, the main information is encoded in the Hamiltonian, which for the purposes of this molecular dynamics simulation is given as $H = T + V$ where the kinetic energy is

$$T = \sum_i \frac{1}{2m_i} p_i^2$$

and the potential energy is given as

$$\begin{aligned} V = & \sum_{\text{CHARMMbonds}} \frac{1}{2} C_{b,\text{CHARMM}} (b - \bar{b})^2 \\ & + \sum_{\text{SCHLICKbonds}} \frac{1}{4} C_{b,\text{SCHLICK}} (b^2 - \bar{b}^2)^2 \\ & + \sum_{\text{angles}} \frac{1}{2} C_a (\cos \vartheta - \cos \vartheta_0)^2 \\ & + \sum_{\text{harm. angles}} \frac{1}{2} C_{a,\text{harm}} (\vartheta - \vartheta_0)^2 \quad (2.1) \\ & + \sum_{\text{dihedrals}} \sum_{k=0}^{k_{\max}} a_k (\cos \omega)^k \\ & + \sum_{i,j \text{ excl}(i,j)=0} \sqrt{V_i V_j} \left(\frac{(r_{0,i} + r_{0,j})^{12}}{r_{ij}^{12}} - 2 \frac{(r_{0,i} + r_{0,j})^6}{r_{ij}^6} \right) \\ & + \sum_{i,j \text{ excl}(i,j)=0} \frac{q_i q_j}{r_{ij}} \end{aligned}$$

2.2.2 Implementation and Parameterisation

The code is set up in a strongly object orientated way, providing classes for atoms, bonds, bond angles, proteins as well as the integrator, the display manager etc. having grown to approximately 9000 lines of C++.

The code contains both potential and force expressions which can all be switched on and off individually and will now be described in turn.

Bond stretch interactions

There are two models for the bond stretch interactions, one taken from the original (1983) CHARMM paper,[17], which is the parametrisation used in [32].

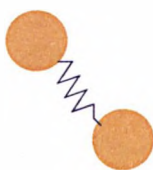


Figure 2.1: Atom Bond

The other model for bond stretch interactions comes from [57], where it is used as a soft constraint to enforce bond length constraints. Its bi-quadratic form reduces the computational cost for force evaluations.

A nonlinear model used by Heyes et al. ([44], [5]) has been used to fit the vibrational frequencies in a testing configuration but is not currently active in the code.

Bond Angle interactions

The bond angle parameters for the trigonometric approximation are given directly in Schlick's article on water clusters, [57].



Figure 2.2: Bond Angle

For the CHARMM parametrisation, their potential expression $k_\theta(\vartheta - \vartheta_0)^2$ is expanded in powers of ϑ and a Taylor series fitted for small deviations from the equilibrium angle to the trigonometric approximation is used, given by Schlick ([54], formula (8.15)) as:

$$C_a \approx 2k_\theta \sin^2 \vartheta_0 \quad (2.2)$$

This approximation reduces computational cost, but Grubmueller and Tavan used the original parametrisation from CHARMM. Both potentials have been implemented, although the harmonic bond angle potential has been tested only briefly.

Dihedral Angle interactions (Torsions)

The dihedral angle is the angle between two planes specified via two adjacent bond angles, see figure 2.3 for an illustration. It is (up to its sign) given by:

$$\cos \omega = \frac{(r_1 \times t_2) \cdot (r_2 \times r_3)}{\|r_1 \times r_2\| \cdot \|r_2 \times r_3\|} \quad (2.3)$$

$$r_1 = q_2 - q_1 \quad (2.4)$$

$$r_2 = q_3 - q_2 \quad (2.5)$$

$$r_3 = q_4 - q_3 \quad (2.6)$$

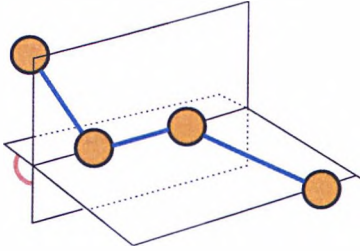


Figure 2.3: Dihedral Angle

Here, $q_1 \dots q_4$ denote the position vectors of the four atoms involved in the dihedral angle. The parameters $a_0 \dots a_5$ are taken from A. Fischer's diploma thesis, [23]. Dihedral angles are used for the simulation of small alkanes which are useful as systems with known metastability. They are *not* used for the Grubmüller-Tavan model of proteins.

Van der Waals interactions

For the alkane simulations, the van der Waals parameters are given by [17]. The interaction exclusion function $\text{excl}(i, j)$ is one whenever two atoms share a common bond or are part of a common bond angle and it is zero otherwise. Computationally, the van der Waals interactions also serve to avoid collisions of oppositely charged particles which would otherwise lead to blow-up in the integrator.

In general, all atom interactions (which are not excluded via $\text{excl}()$) are computed, yielding an $\mathcal{O}(N^2)$ algorithm. A nearest-neighbour boxing strategy is implemented and has been tested, but as the advantage only becomes pal-



Figure 2.4: Van der Waals Interaction

table at large (> 100) numbers of atoms or in the case of well-separated atoms, this further approximation is not made here. Even the more advanced strategies like particle mesh Ewald methods (which I have considered for implementation and studied in some detail) or Multipole methods are reported (e.g. in [47]) only to show significant advantages beyond 100 to 1000 atoms in the simulation.

Electrostatic Interactions

For each atom, one can impose an electrostatic charge q_i . In the case of water, charges to reproduce qualitative behaviour of water droplets are well-known (reported in mutual agreement by [57] and [19], Table 1, SPC and F3C),



whereas in the case of the charges given by Grubmüller and Tavan in [32] the only information available until the end of coding work on the program was their Fig. 1, which was qualitatively approximated by two periods of a cosine with amplitude $0.5e$. See below for a further discussion of this point. For the protein simulations, interactions between partial charges are only allowed if the atoms do not share a bond or bond angle. This is as specified in [17].

Figure 2.5: Electrostatic Coulomb Interaction

2.2.3 Verification of Force expressions

The Force induced by the potentials is given via

$$F = -\nabla V. \quad (2.7)$$

This is amenable to direct numerical verification by numerical differentiation of the potential. In view of the rather involved force expressions, especially for the dihedral angle terms, this turns out to be a very valuable tool.

The approach to verify the implemented force expression adopted here is to compute a numerical approximation of the gradient of the potential in a component-wise fashion:

$$\bar{F}_i(x) = -\frac{V(x + he_i) - V(x)}{h} \quad i \in \{1, \dots, N\} \quad (2.8)$$

Then the deviation from the implemented force is computed, and the two-norm of the

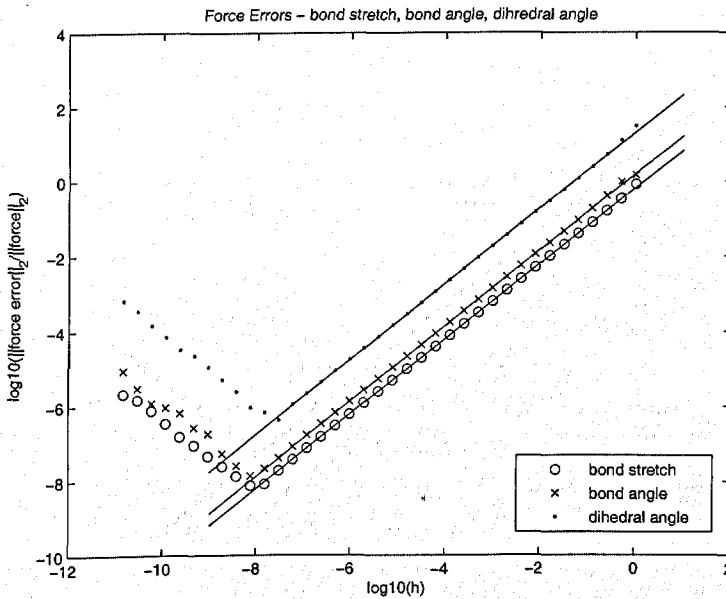


Figure 2.6: Force errors: BS, BA, DA

relative error is output:

$$E(h) = \frac{\|\tilde{F} - F\|_2}{\|F\|_2} \quad (2.9)$$

Of course, this is still a function of the position vector x , so an initial condition x is constructed by placing all atoms of the protein under test at random in a box of size $2 \times 2 \times 2a^3$. In a doubly logarithmic plot, the force error incurred is plotted as a function of h , yielding plot 2.6. This was obtained for a butane atom configuration. The slope of the fitted lines are 1.0002, 1.0070 and 1.0068 for the CHARMM bond-stretch, trigonometric bond-angle and dihedral angle interaction respectively. For the van der Waals and the electrostatic interaction a cluster of four water molecules with random initial conditions produced in the same way was used. The plot obtained is given in figure 2.7 where the slopes of the fitted lines are 1.0091 and 1.0058 for the van der Waals and the electrostatic interaction respectively.

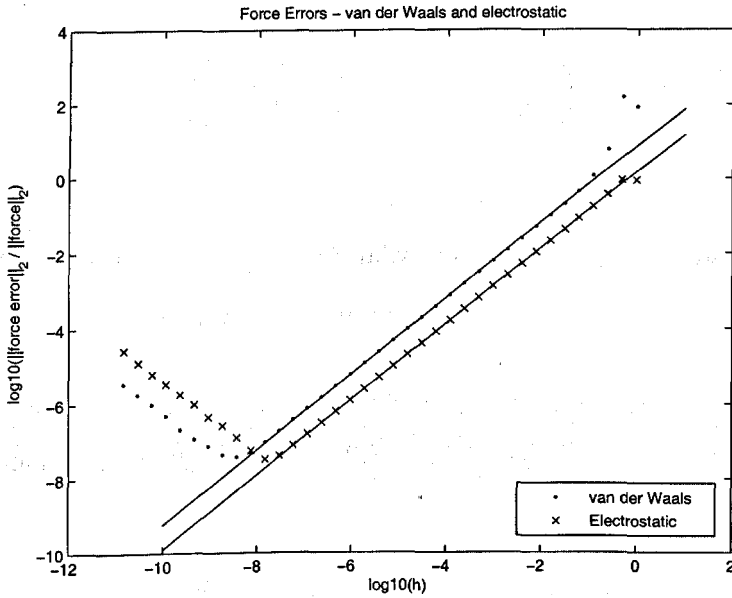


Figure 2.7: Force errors: vdW, Coulomb

2.2.4 The Integrator

The scheme used to integrate the Hamiltonian equations of motion is

$$\begin{aligned} q_{n+\frac{1}{2}} &= q_{n-\frac{1}{2}} + \Delta t M^{-1} p_n \\ p_{n+1} &= p_n + \Delta t F(q_{n+\frac{1}{2}}) \end{aligned} \quad (2.10)$$

where $M = \text{diag}(m_i)$ is the mass matrix. If a starting step of the form

$$p_1 = p_{\frac{1}{2}} + \Delta t F(q_{\frac{1}{2}})$$

is used, then starting from an initial condition $q_{\frac{1}{2}}, p_{\frac{1}{2}}$ the method is of order 2 and is sometimes referred to as the Störmer/Verlet integrator, e.g. in [54]. It belongs to the class of symplectic integrators which, up to floating point rounding error, constitute a symplectic transformation in every step of integration. A full account of the theory can be found in [14]. One of the essential features of these integrators is that they exactly (up to floating point accuracy) preserve phase space volume. They also approximately conserve a Hamiltonian that is close to the true Hamiltonian.

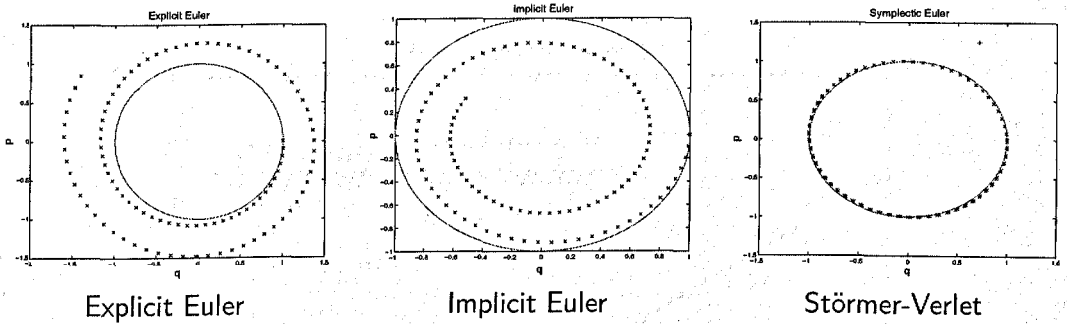


Figure 2.8: Comparison of Integrators

To illustrate typical behaviour of this integrator, consider the case of a simple harmonic oscillator given by

$$\begin{aligned}\dot{q} &= p \\ \dot{p} &= -q.\end{aligned}$$

Using an explicit Euler scheme to integrate these equations yields the trajectory given on the left of figure 2.8, whereas an implicit Euler scheme yields the plot in the middle of 2.8. In both these, it is clear that phase space volume will be either gained or lost. Using the above Störmer-Verlet scheme yields the plot given in figure 2.8 where the conservation of phase space volume is mirrored by the fact that the displayed ellipse has the same volume as the circle (which corresponds to the true solution). Note, however, that the distance from the origin (whose square corresponds to the energy) is not exactly constant, so there is no conservation of energy. The long term properties of these integrators have seen renewed interest recently.

The order of the numerical scheme (2.10) in the current implementation is verified in the following section.

Order of convergence

A diatomic molecule with one CHARMM bond is started with initial positions in the energy minimum position and velocities such that speeds and positions rise to order unity size. The bond-stretch interaction is the only active interaction in this model. For a fixed final time $T_F = 1$, a variety of timesteps is used, comparing the simulation result

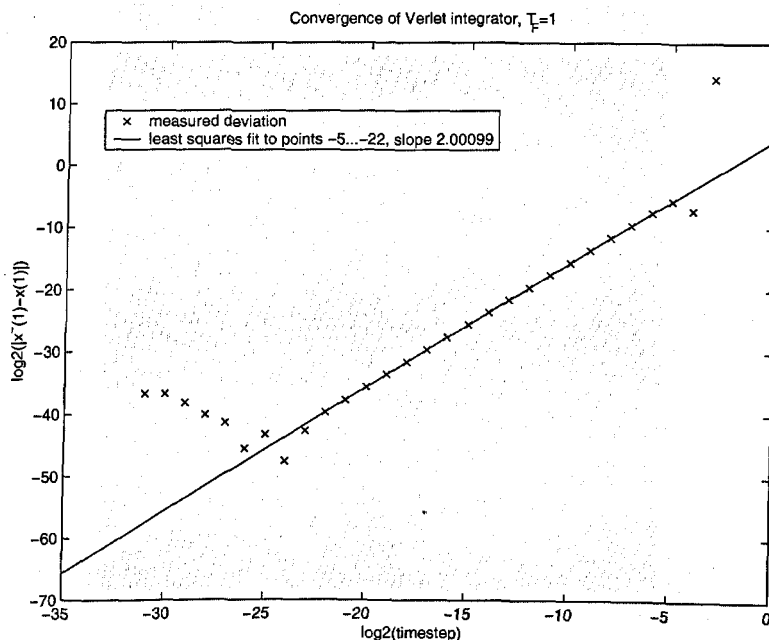


Figure 2.9: Verlet Integrator Verification

at the final time to the analytic result. A doubly logarithmic plot reveals an intercept with rounding errors around a timestep of size $\Delta t \approx 2.4 \cdot 10^{-7}$. The slope of the least squares fitted line is -2.00099 , corroborating the method being of order two. Note, however, that the precise value of the slope is easily modifiable by including or excluding some of the extreme points.

Order of convergence 2

A typical 100 residue initial condition with cosine charge distribution and in all other aspects following [32], is evolved (approximately conserving energy) until a collapsed condition is obtained, see the figure below.

This is used as a starting condition for simulation for 49fs, using timesteps from $49 \cdot 2^{-4} \dots 49 \cdot 2^{-13}$ fs. Note that this simulation takes place at high total kinetic energies (of the order of $4000 \frac{\text{kcal}}{\text{mol}}$). The result obtained for the finest timestep is considered a close approximation to the true solution and the 2-norm deviations of the spatial coordinates for the fixed final time and varying timestep is plotted in a doubly logarithmic

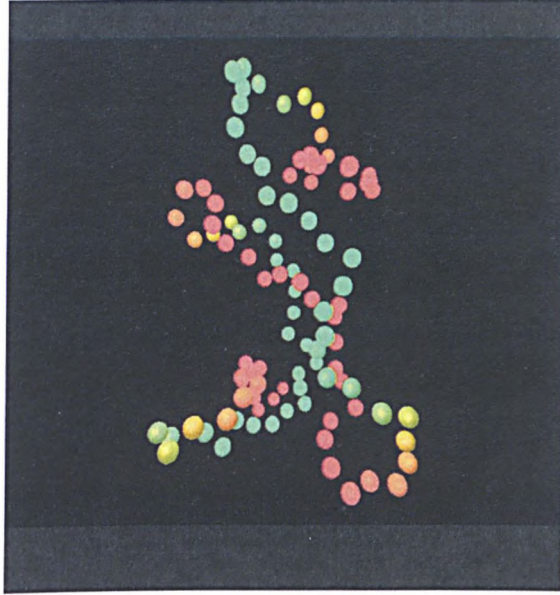


Figure 2.10: Semifolded intermediary configuration

plot. Again, a line is fitted using a least squares fit and the slope obtained is again close to two, see figure 2.11.

Energy conservation

To test the potential evaluation functions (and not only the force evaluation), the same semi-folded initial condition as above (2.10) is used as a starting condition. The potential energy terms for the four contributions (using CHARMM bonds and trigonometric angles, only) are exported from the program as well as spatial coordinates. The corresponding velocities are computed in a post-processing step from the coordinates to circumvent the problem of split-timesteps. (Having x_n but $v_{n+\frac{1}{2}}$ and $v_{n-\frac{1}{2}}$ in the program...) The total energy of the protein as a function of time is plotted and its standard deviation from its equilibrium value is computed.

Plotting the RMS deviation of the total energy from its average value as a function of timestep in a doubly logarithmic plot yields the plot in figure 2.12. A line is fitted to the data-points using least squares (and omitting the leftmost data-point as this is meant to elucidate the asymptotic behaviour only) and its slope is found to be

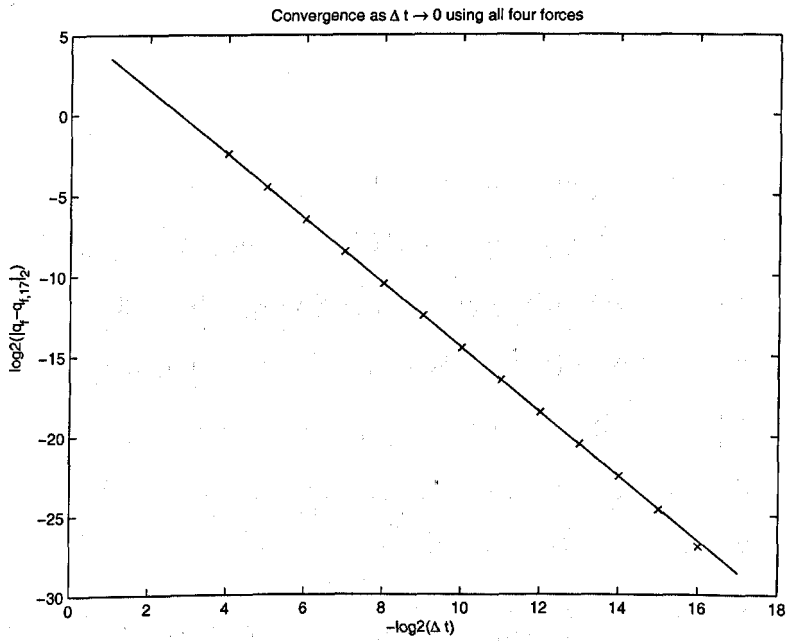


Figure 2.11: Verlet Integrator Verification 2

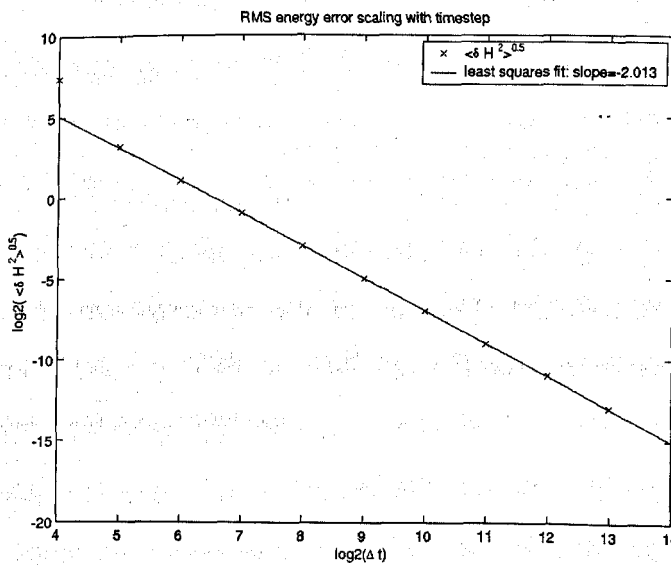


Figure 2.12: RMS Energy deviations

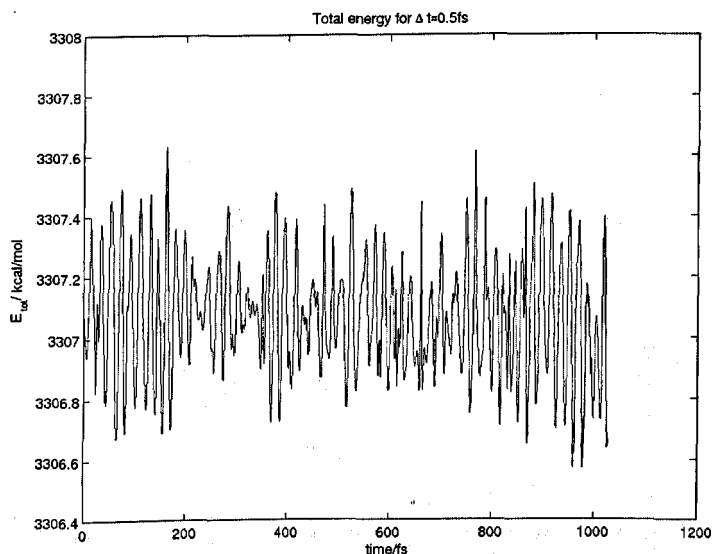


Figure 2.13: Total Energy fluctuations

-2.013 , which is close to two, as expected from [47], p.83.

A plot of the total energy as a function of time at these high temperatures using a timestep typical of later simulations can be seen in figure 2.12.

2.2.5 Lennard-Jonesium

Another qualitative check was done using periodic boundary conditions on a van der Waals liquid (using parameters for Argon from [60]). The observed statistical quantity is the radial distribution function, $g(r)$, a histogram of which is given in figure 2.14.

Good qualitative agreement with the plot in [47] is obtained. As quantitative data for these plots (at $T = 300K$, $\rho = 1350\text{kg/m}^3$) was not immediately available, a quantitative evaluation was renounced.

2.2.6 Water

In order to verify the interpretation of the parameters, a simple, well-studied molecule was needed. Using [57], a parametrisation specifically for water is analysed in the sequel. Note that the above article uses a different potential for the bond-stretching term for

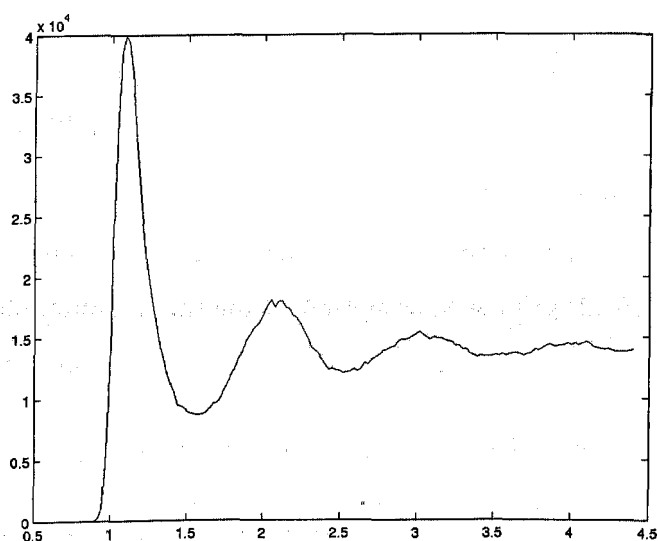


Figure 2.14: Radial Distribution function for Argon

reasons of ease of computations. Schlick et al. restrict the Coulomb interaction to act intermolecularly only (as opposed to both, intramolecularly and intermolecularly). Furthermore, they have a van der Waals interaction only between oxygen atoms belonging to different water molecules, i.e. Hydrogen atoms are completely excluded from van der Waals interactions. To compare with the simulations in that paper, I try to reproduce some of their TABLE 1 here (all energies in units of $\frac{\text{kcal}}{\text{mol}}$):

# Molecules	E_{tot}	E_{coul}	E_{vdW}	E_{bond}	E_{angle}
1	6.82385e-31	0	0	2.59956e-31	4.22429e-31
2	-6.9391	-9.43512	2.1476	0.258041	0.0903812
4	-32.4175	-46.0538	10.8098	2.02165	0.804886

Note that the energy minimisations are done using steepest descent with a linear search strategy guaranteeing monotonicity. There is no guarantee to find the absolute minimum, however, and there seem to be quite a few local minima.

Comparing to the data given in [57], the agreement is found to be approximate rather than exact. As a strong dependence on parameters (e.g. Coulomb charges) has been observed, it is speculated that small differences in conversions (e.g. of charge

units, in particular in view of the limited number of digits reported for charge conversion in [57]) may greatly affect minimal energy conformations.

2.2.7 Metastability

Using CHARMM bond-stretch, trigonometric bond angle, standard van der Waals and electrostatic interaction on a linear chain of 100 CH₂ extended atoms, just as in Figure 2.10, long-time simulations of the protein have been done using the following simulation protocol specified in [32]:

1. Simulate starting from a stretched configuration for 2ns, using 1fs timesteps.
2. Cool down by velocity rescaling every 10th timestep such that at every tenth timestep the kinetic temperature of the protein is the desired 300K.
3. Observe for a further 2ns, verifying that kinetic temperature remains around 300K.

The initial configuration phase was deemed necessary by Grubmüller and Tavan to ensure "proper exploration of phase space", but from the simulation results, it does not seem to bring about much of a change after the first 200ps.

Observation in step 3 above shows that the protein has indeed cooled down and that the potential and kinetic energy degrees of freedom have equilibrated. This 2ns period together with some extra observation time (usually 5ns) was also used to look for metastability, which was expected to occur at a rate of $1.89ns^{-1}$. No such rare transitions of the magnitude reported in [32] have been found.

The potential reasons for this may be:

- Using trigonometric angles instead of harmonic angles makes a large difference in overall qualitative behaviour even though the force errors at the angle deviations observed at 300K are only a few per cent.
- Most importantly, the charge distribution is critical and the cosine fit to it destroys qualitative features of the dynamics of this particular protein model.
- Not enough phase space volume has been scanned.

Pursuing Grubmüller's model further even though no metastabilities have been observed so far might not seem desirable. While the precise charge distribution as well as the original code have now been made available to me, reverse engineering this code did not seem conducive to good research as large amounts of coding would still be required and the protein model is, after all, just a simplistic model. The main focus of the present chapter is to gain some understanding and intuition for molecular dynamics as well as to provide a trusted source of molecular dynamics data for which Butane seemed a suitable example. This will be described in the next section.

2.3 Butane

A very simple organic molecule exhibiting metastability at room temperature is butane. It has been the subject of A. Fischer's diploma thesis ([23]) on a hybrid Monte Carlo method precisely for this reason. Doing butane necessitated implementing dihedral angles which required some debugging effort so careful verification was in order. A typical trajectory at three orders of magnitude of time resolution can be seen in figure 2.15.

It is easy to verify that dihedral angle potential and force expression fit each other, but more verification can be done exploiting theorem 1 in [23], i.e. sampling dihedral angles from the canonical ensemble for the butane model yields exactly the same sample statistics for dihedral angles as sampling for the canonical measure $\exp\left(\frac{-H(\omega)}{kT}\right)$ for the dihedral angle alone.

In order to do this, the Generalised Verlet Algorithm for Langevin Dynamics (p.437 in [54]) for (approximately) sampling from the canonical ensemble was implemented. Figure 2.16 shows good agreement between the histogram and the expected probabilities.

2.4 Conclusion

The explorations of molecular dynamics described above have provided a learning field for molecular dynamics and high performance coding as well as a medium size C++

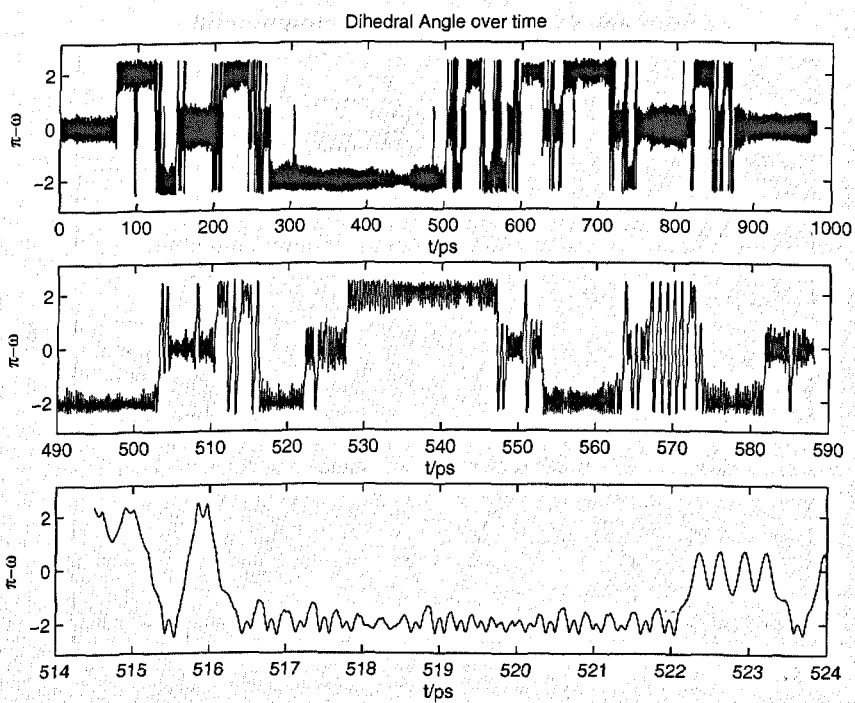


Figure 2.15: Butane Dihedral Angle Trajectory

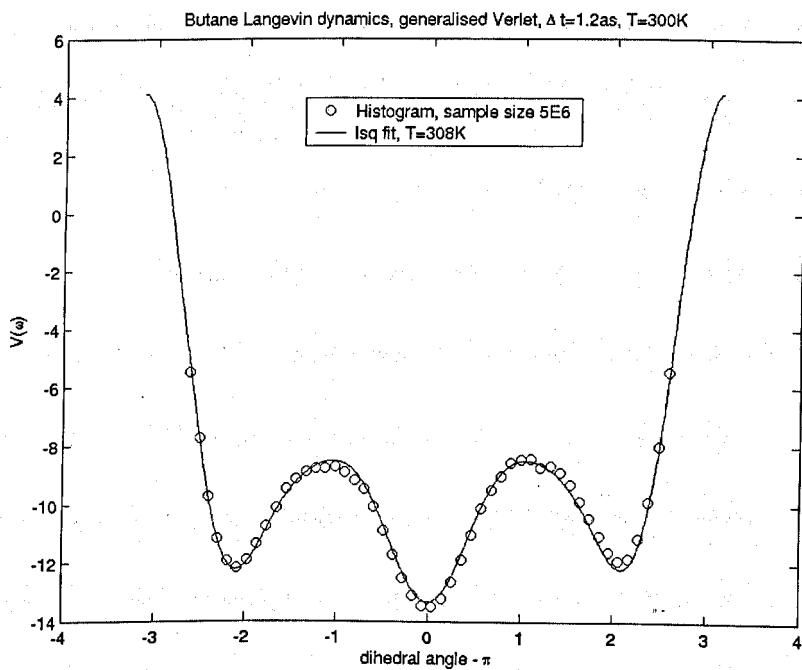


Figure 2.16: Langevin Dynamics for Butane – Invariant Measure and Empirical Density

code for MD simulations and energy optimisations. This code has been verified in a multitude of ways and will thus be viewed as a sufficiently reliable source of molecular dynamics trajectories. These trajectories will be used to test the fitting algorithms to be described in the next two chapters.

Further avenues of research into molecular dynamics might include a study of Alanine dipeptide in aqueous solution, which seems to be a standard example of conformational metastability, possibly using GROMACS to deal with the explicit solvent representation. While this might take the theoretical work closer to actual application, in view of the intricacies of long trajectory simulation I have chosen to follow most applied mathematicians' approach and concentrated on developing mathematical and statistical tools using a toy example.

Chapter 3

First Order Stochastic Differential Equations

The easiest way to explain this idea is to contrast it, for example, with advertising. Last night I heard that Wesson oil doesn't soak through food. Well, that's true. It's not dishonest; but the thing I'm talking about is not just a matter of not being dishonest; it's a matter of scientific integrity, which is another level. The fact that should be added to that advertising statement is that no oils soak through food, if operated at a certain temperature. If operated at another temperature, they all will – including Wesson oil. So it's the implication which has been conveyed, not the fact, which is true, and the difference is what we have to deal with.

R.P. Feynman, Caltech Commencement address 1974

3.1 Introduction

After briefly reviewing standard results about parameter estimation using maximum likelihood methods for non-degenerate 1D diffusion processes in continuous time in section 3.2, simple implementations of these estimators for discrete time are considered in section 3.3. Random number generators are discussed briefly and parameter estimation for a family of SDEs is considered. The algorithm obtained here is applied to time

series from Langevin dynamics for a single Butane molecule in section 3.4 and problems inherent in fitting SDEs with rough paths with finite quadratic variation to data from smooth Hamiltonian systems are highlighted.

3.2 Continuous Time Path fitting

This chapter will deal with fitting drift and diffusion parameters in stochastic differential equations of the form

$$dx = \sum_{i=1}^c \theta_i f_i(x) dt + \sigma dB \quad x(0) = x_0 \quad (3.1)$$

where f_i are suitably well-behaved (e.g. globally Lipschitz) force functions, x_0 is a deterministic starting condition and B is standard Brownian motion where $\sigma \in \mathbb{R}^+$ is the diffusivity constant, $\theta_i \in \mathbb{R}$ are the drift parameters and $c \in \mathbb{N}$ is the number of force terms to be used. Given a piece of trajectory, $\{x_s\}_{s \in [0, T]}$, the parametric estimation problem is how to estimate the diffusion coefficient σ and the drift coefficients θ_i . Estimating the diffusion coefficient is straightforward under these conditions but there are many methods for estimating drift parameters. I will consider maximum likelihood estimators as the theoretical understanding of these estimators is well-developed and they are frequently applied in actual practise. Also, in this context, they are normally easy to generalise to a Bayesian framework which might be useful for applications in molecular dynamics where posterior variances as well as expected values are of interest.

3.2.1 Estimating Diffusivity

In general, given a continuous piece of trajectory, however short, estimating the diffusivity is considered easy, even when multiplicative noise is present. In the case of the process given here, (3.1), a few remarks will show how σ can be estimated using only modestly technical results.

Firstly, (3.1) can be written in integral form (which is its very definition):

$$x(t) - x(0) = \int_0^t \sum_{i=1}^c \theta_i f_i(x(s)) ds + \sigma B(t) \quad (3.2)$$

Now, the solution of this SDE is a continuous semimartingale with

$$A_t = \int_0^t \sum_{i=1}^c \theta_i f_i(x(s)) ds$$

being continuous, adapted and of locally bounded variation and $M_t = \sigma B(t)$ being a (local) martingale so that overall

$$x(t) - x(0) = A_t + M_t$$

holds.

Using theorem (8.6) of [13] it is clear that the approximate quadratic variation converges to the true quadratic variation under mesh refinement, where the approximate quadratic variation of x is given as follows:

$$Q_t^\Delta(x) = (x(t) - x(t_K))^2 + \sum_{k=1}^K (x(t_k) - x(t_{k-1}))^2$$

Here, $\Delta = \{t_0 = 0 < t_1 < t_2 \dots < t_K \leq t\}$ is a mesh of the interval $[0, t]$. If a sequence of meshes, Δ_n goes to infinite refinement in the sense $\lim_{n \rightarrow \infty} \min\{t - t_K^n, \min\{t_{i+1}^n - t_i^n\}\} = 0$ then the theorem's statement implies that

$$Q_t^{\Delta_n}(x, x) \longrightarrow \sigma^2 \langle B(t), B(t) \rangle.$$

The resulting estimator of the diffusivity parameter σ is thus given as

$$\hat{\sigma}^2 = \frac{1}{t} \lim_{n \rightarrow \infty} Q_t^{\Delta_n}.$$

In particular, the sequence of meshes can be chosen uniformly, i.e. just using $t_k^{(n)} = t \cdot \frac{k}{n}$, so that the resulting estimator is finally given as

$$\hat{\sigma}^2 = \frac{1}{t} \lim_{n \rightarrow \infty} \sum_{i=1}^n \left(x\left(\frac{i}{n}t\right) - x\left(\frac{i-1}{n}t\right) \right)^2 \quad (3.3)$$

Other ways of estimating the diffusivity applied mostly by practitioners from physics and chemistry include fitting the invariant density or the Laplace transformation of the autocorrelation provided the drift parameters are known. This will be covered in more detail in chapter 5.

3.2.2 Estimating Drift Parameters

Drift estimation in continuous time is a harder problem, as the SDE in question is required not only to have a unique strong solution but also to possess certain ergodic properties. Sufficient conditions specialised to the SDE in question, (3.1), will be shown here. All conditions and theorems quoted in this subsection are taken from [62] or [13].

Firstly, assuming the f_i are locally bounded and measurable, existence of a unique weak solution is guaranteed by theorem 5.3.2 of [13] if

$$x \sum_i \theta_i f_i(x) \leq A(1 + x^2) \quad (3.4)$$

for some $A > 0$. This is implied by Kutoyants' condition to ensure ergodicity, $\mathcal{A}_0(\Theta)$. It is assumed that only certain vectors $(\theta_1, \dots, \theta_c)^T \in \mathbb{R}^c$ from an open bounded subset $\Theta \subset \mathbb{R}^c$ are admissible. For these drift parameters the assumption $\mathcal{A}_0(\Theta)$ is:

$$\forall(\theta_i) \in \Theta : \lim_{|x| \rightarrow \infty} \text{sgn}(x) \sum_{i=1}^c \theta_i f_i(x) < 0 \quad (3.5)$$

To ensure identifiability the information matrix

$$I(\theta) = \mathbb{E}_\theta (f_i(\cdot) f_j(\cdot))_{i,j \in \{1, \dots, c\}},$$

where the expectation is taken with respect to the induced invariant measure of the SDE, must be positive definite uniformly on compact subsets $\mathbb{K} \subset \Theta$ of parameter space:

$$\inf_{\theta \in \mathbb{K}} \inf_{|e|=1} e^T I(\theta) e > 0 \quad (3.6)$$

This is sufficient to ensure condition \mathcal{A} on p.115/116 of [62] holds. Now, theorem 2.8 from [62] can be used to infer the following:

Theorem 1. *Let conditions (3.6) and (3.5) hold. Then for any fixed $\theta \in \Theta$, any of the $\hat{\theta}_T$ attaining the supremum over Θ of the Radon-Nikodym derivative*

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}} (\theta; \{x(s)\}_{s \in [0, T]}) \quad (3.7)$$

of the measure on path space induced by θ w.r.t. to the measure induced by $\theta = 0$ is asymptotically unbiased:

$$\lim_{T \rightarrow \infty} \mathbb{E}_\theta (\hat{\theta}_T - \theta) = 0 \quad (3.8)$$

Note that the supremum might be attained on the boundary of Θ , i.e. $\hat{\theta}_T \in \partial\Theta$ has positive probability in general. It should be stressed that the aforementioned theorem states far more than this, including asymptotic consistency, asymptotic normality, convergence of higher moments and asymptotic efficiency, all uniformly on compact subsets of parameter space. Since this chapter will be concerned with the behaviour of the estimator's bias, $\mathbb{E}(\hat{\theta} - \theta)$, this version seemed most appropriate.

In the setup (3.1), the maximum likelihood estimator is actually unique and can be given explicitly. It suffices to note that the Radon-Nikodym derivative (3.7) is given explicitly as

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}}(\theta) = \exp\left(\frac{1}{\sigma^2} \int_0^T \sum_{i=1}^c \theta_i f_i(x(t)) dx_t - \frac{1}{2\sigma^2} \int_0^T \left[\sum_{i=1}^c \theta_i f_i(x(t)) \right]^2 dt\right)$$

Taking partial derivatives with respect to the θ_i and equating them to zero results in the following drift estimator:

$$\hat{\theta} = \left(\int_0^T f_i(x(t)) f_j(x(t)) dt \right)_{i,j \in \{1, \dots, c\}}^{-1} \begin{pmatrix} \int_0^T f_1(x(t)) dx_t \\ \vdots \\ \int_0^T f_c(x(t)) dx_t \end{pmatrix} \quad (3.9)$$

3.3 Numerical Implementation

In order to put the estimators (3.3) and (3.9) into practise, the problem has to be discretised. After introducing two perspectives on estimation in the discrete time framework, this section will detail the numerical implementation of these discretised estimators, highlighting issues concerning the random number generator and numerically examine convergence as the discretisation timestep tends to zero.

Assuming that observations $\{x_i\}_{i \in \{1, \dots, N\}}$ at equispaced timepoints $t_{ii \in \{1, \dots, N\}}$ with spacing Δt of the process (3.1) are given, the task is to estimate the parameters θ_i and σ .

While results implying asymptotic consistency are available in the limiting case $N\Delta t \rightarrow \infty$, $\Delta t \rightarrow 0$ from [50], the interest in this chapter is to arrive at practical implementations of estimators, quantify errors as a function of Δt and apply these

estimators to the practical problem at hand. The approach taken is therefore to develop these estimators using concrete numerically worked examples.

One perspective that can be taken in constructing estimators in the discrete time framework is to approximate the sums and integrals occurring in (3.3) and (3.9) by Riemann sums. In the case of diffusion estimation, it is straightforward to take the available data, insert it into (3.3) and evaluate for some finite Δt . In the case of drift estimation, the integrals occurring in (3.9) can be replaced by Riemann sums so that the resulting estimator reads as follows:

$$\hat{\theta} = \left(\sum_{p=1}^N f_i(x_p) f_j(x_p) \Delta t \right)_{i,j \in \{1, \dots, c\}}^{-1} \begin{pmatrix} \sum_{p=0}^{N-1} f_1(x_p) (x_{p+1} - x_p) \\ \vdots \\ \sum_{p=0}^{N-1} f_c(x_p) (x_{p+1} - x_p) \end{pmatrix} \quad (3.10)$$

Another perspective is obtained by replacing the stochastic differential equation (3.1) by a stochastic difference equation

$$x_{p+1} = x_p + \Delta t \sum_{i=1}^c \theta_i f_i(x_p) \Delta t + \sqrt{\Delta t} \sigma \xi_p \quad (3.11)$$

where $\xi_p \sim \mathcal{N}(0, 1)$ are i.i.d. normal random variables. This difference equation will be referred to as a statistical model for the diffusion process. One can then employ a maximum likelihood approach yielding exactly the same MLE as (3.10), however this can be used within a Bayesian framework. Rather than using an approximating difference equation, the exact transition density could be used, however, depending on the particular choice of force functions f_i , this is not normally available analytically, so one can resort to statistical models instead. Other approaches involve perfect inference and inference via particle methods.

3.3.1 Diffusion Coefficient

As a means of introducing the numerical implementation of estimators for the problem (3.1), the 1D Ornstein-Uhlenbeck process is considered as a simple first example:

$$dx = -\alpha x dt + \sigma dB \quad (3.12)$$

Here, we assume $\alpha, \sigma \in \mathbb{R}_+$. Given observations $x_i, i = 0, \dots, N$ at times $t_i = i\Delta t$ for some $\Delta t > 0$, the aim is to provide an estimator for the parameters α and σ .

As explained in the previous subsection, one approach is to formulate a statistical model which is in some sense a discrete version of (3.12):

$$x_{n+1} = x_n - \Delta t \alpha x_n + \sigma \sqrt{\Delta t} \xi_n$$

Here, $\xi_n \sim \mathcal{N}(0, 1)$ are independent identically distributed normal random variables.

Based on the quadratic variation of paths from (3.12), the estimator (3.3) is adapted as follows:

$$\hat{\sigma}^2 = \frac{1}{N\Delta t} \sum_{n=0}^{N-1} (x_{n+1} - x_n)^2 \quad (3.13)$$

Asymptotic consistency of this estimator is assured as set out in subsection 3.2.1. To assess the order of error, one can use an Ito-Taylor expansion for the process (3.12) at time t_n to tackle the term $(x_{n+1} - x_n)$:

$$\begin{aligned} \mathbb{E} \left[\frac{(x_{n+1} - x_n)^2}{\Delta t} \right] &= \mathbb{E} \left[\frac{-\alpha x_n \Delta t + \xi_n \sqrt{\Delta t} + \mathcal{O}(\Delta t^{1.5})}{\Delta t} \right] \\ &= \sigma^2 + \Delta t \mathbb{E} [\alpha^2 x_n^2] + \mathcal{O}(\Delta t^2) \end{aligned}$$

(More thorough consideration is given to Ito-Taylor expansions in the next chapter where higher orders of accuracy are required.) The Ito-Taylor expansion shows the order of error but bounding the error terms uniformly in time and over state space may be hard or impossible, depending on the particular process at hand.

3.3.2 Numerical Validation: Pitfalls

In order to verify the C++-code used for fitting coefficients the above estimator for σ , (3.13), has been implemented. While testing the code, problems concerning random number generators have been observed and these will be highlighted in statistically significant experiments.

To generate paths for the experiments, a final time of $T_f = 4000$ is used and the parameters for the Ornstein-Uhlenbeck process are

$$\sigma = 1 \quad \alpha = 1.$$

While it is possible to generate sample paths satisfying the exact statistics for (3.12) by just choosing a Gaussian with appropriate mean, variance and correlation at each timestep t_i , samples are generated here using a subsampled Euler-Maruyama method.

Given a data point x_i , the next data point in a path, x_{i+1} , is generated using several steps of the Euler-Maruyama algorithm. The number of intermediary steps, k , is called the subsampling factor. Using the notation $x_i^{(j)}$ for the j -th intermediary step involved in generating the $(i+1)$ th sample point from the i th sample point, this can be written as follows:

$$\begin{aligned} x_i^{(0)} &= x_i \\ x_i^{(j+1)} &= x_i^{(j)} - \frac{\Delta t}{k} \alpha x_i^{(j)} + \sqrt{\frac{\Delta t}{k}} \sigma \xi_i^j \\ x_{i+1} &= x_i^{(k)} \end{aligned} \tag{3.14}$$

For the subsampling factor $k = 1$, the statistical model (3.13) and the generation of the data (3.14) coincide, so it is expected that the estimate for σ will be exact and this is indeed what is observed.

As k is increased, the sample x_{i+1} follows the exact statistics more and more. More precisely, one observes that the Euler-Maruyama method converges weakly for this SDE so that, in particular, first and second moments converge to the correct values as $k \rightarrow \infty$. Since all increments are Gaussian, it is clear that the quantity $x_{i+1} - x_i$ occurring in the estimator (3.13) is approximated correctly.

Using a subsampling factor of $k = 30$, the simulations are repeated with the above parameters, averaging the results over $N_s = 100$ runs with different random seeds to both control and estimate Monte Carlo sampling error.

It is clear from the plots given in figures 3.1 and 3.2 that the estimator $\sqrt{\sigma^2}$ does not converge to σ as $\Delta t \rightarrow 0$. From the second plot, (3.2), it can be seen that this is not attributable to Monte Carlo sampling error.

After considerable simplification of the code, the only potential culprit left is the random number generator. Linear congruence generators are known to be usable for evaluating integrals using Monte Carlo Methods if the correct parameters are chosen. But do they perform well for the 'differential' task at hand?

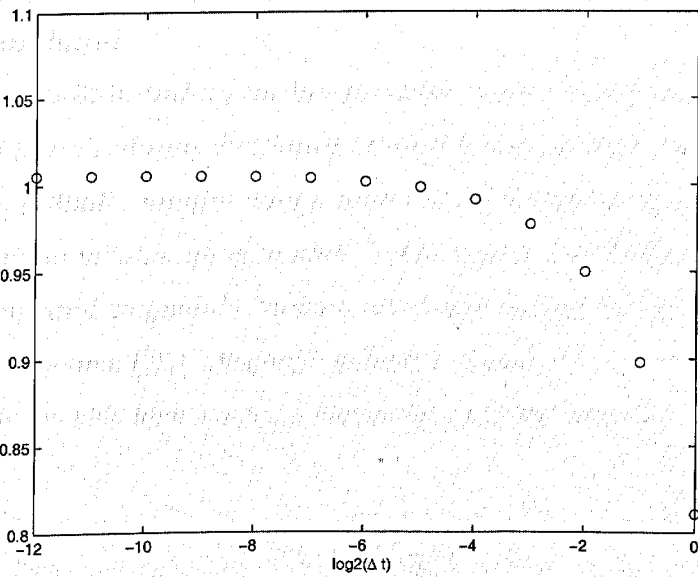


Figure 3.1: σ Sampling Error

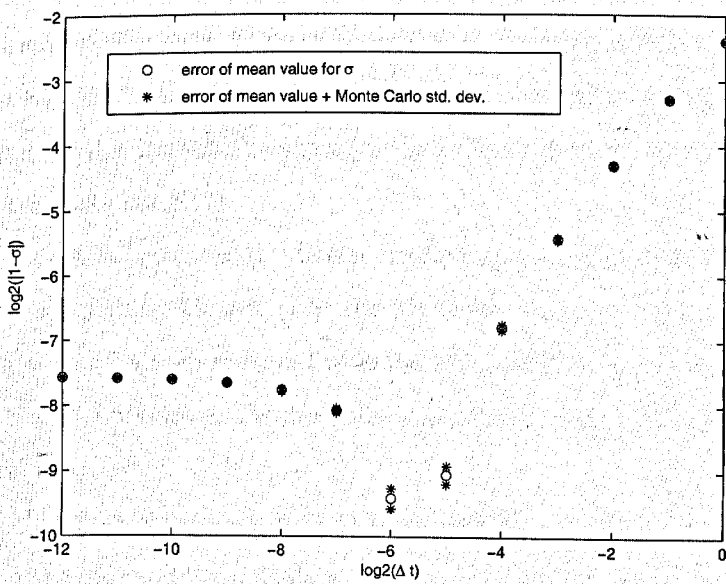


Figure 3.2: σ Sampling Error - logarithmic

In order to demonstrate the incorrect statistical behaviour of the standard linear congruence generator in the GNU c-library `glibc`, version 2.2.5-34, the following experiment is conducted:

Using the built-in random number generator, pseudorandom integers generated by `rand()` which are uniformly distributed between 0 and `RANDMAX` are generated. Dividing these by `RANDMAX`, samples from a uniformly distributed (on a grid with spacing $\frac{1}{\text{RANDMAX}}$) random variable are generated. These quasi-uniformly distributed random numbers are then used to generate normally distributed random samples employing the Box-Müller scheme from ([21], chapter 7, section 2, p.216).

Using k samples from normally distributed random variables, n_i , the random variable

$$s = \sqrt{\frac{1}{k}} \sum_{i=1}^k n_i \quad (3.15)$$

is sampled repeatedly. The random variable s is, of course, itself a Gaussian with mean zero and variance one. Using $N_s \approx 2 \cdot 10^7$ Monte Carlo samples of s , the observed variance for $k = 1, \dots, 49$ is given in figure 3.3

It can be seen from this figure that the deviation at $k = 30$ cannot be attributed to Monte Carlo sampling error. It is not entirely clear, however, whether this error can be attributed to `RANDMAX` being finite, i.e. the fact that the uniformly distributed random variables input to the Box-Müller procedure are distributed on an equispaced grid in $[0, 1]$ rather than all over $[0, 1]$.

In any case, it is clear that a better random number generator is required. Consulting the Gnu Scientific Library, [18], the Mersenne Twister is recommended, citing [46]. The experiment is then repeated with the Mersenne twister yielding plot (3.4). It can be seen that the Mersenne Twister combined with the Box-Müller procedure passes this statistical test.

Note that the dotted lines in (3.4) represent the $1\text{-}\sigma$ limits for the average over N_s realisations of the random variable s , so that roughly $1/3$ of the points is expected to lie outside the bounds.

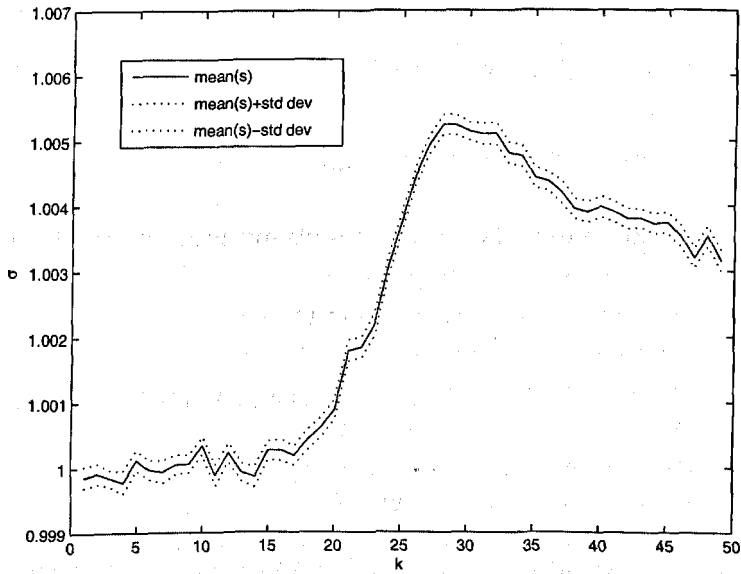


Figure 3.3: Variance Error for LNC RNG

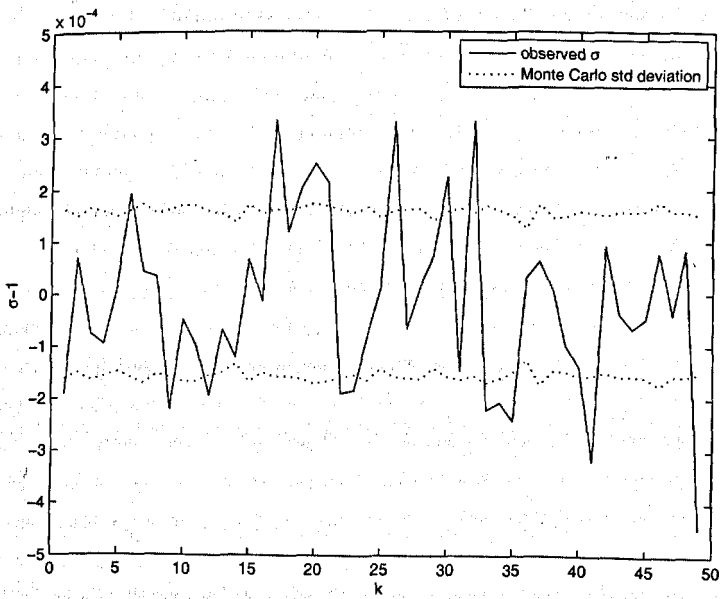


Figure 3.4: Mersenne Twister – Variance

3.3.3 Drift Parameters: Polynomial Potential

In order to move towards applicable parameter estimation procedures, a first order SDE with constant diffusion coefficient and variable force expression is considered:

$$dx = \sum_{i=1}^c \theta_i f_i(x) dt + \sigma dB \quad (3.16)$$

Here, the force basis functions are chosen simply to be powers of x

$$f_i(x) = x^i, \quad (3.17)$$

and the potentials are defined as follows:

$$V_i(x) = \int_0^x f_i(y) dy$$

An arbitrary additive constant can be chosen for the potentials, which has been fixed here by starting the integral at 0.

Using (3.10) as an estimator for the drift parameters θ_i , the following abbreviations are introduced:

$$M_{i,j} = \sum_{n=1}^N \Delta t f_i(x_n) f_j(x_n) \quad (3.18)$$

$$b_i = \sum_{n=0}^{N-1} f_i(x_n) (x_{n+1} - x_n). \quad (3.19)$$

The estimator (3.10) can now simply be written as

$$\hat{\theta} = M^{-1}b \quad (3.20)$$

An analysis of truncation error incurred using the statistical model can be performed using Ito-Taylor expansions and predicts a bias of order $\mathcal{O}(\Delta t)$ for the estimator (3.20).

The estimator (3.20) has been put into practise and the implementation is tested using the following parameters for an example:

$$\theta_1 = 4 \quad \theta_2 = -0.3 \quad \theta_3 = -4 \quad \theta_4 = 0 \quad \theta_5 = 0 \quad \sigma = 0.8 \quad N_s = 1000$$

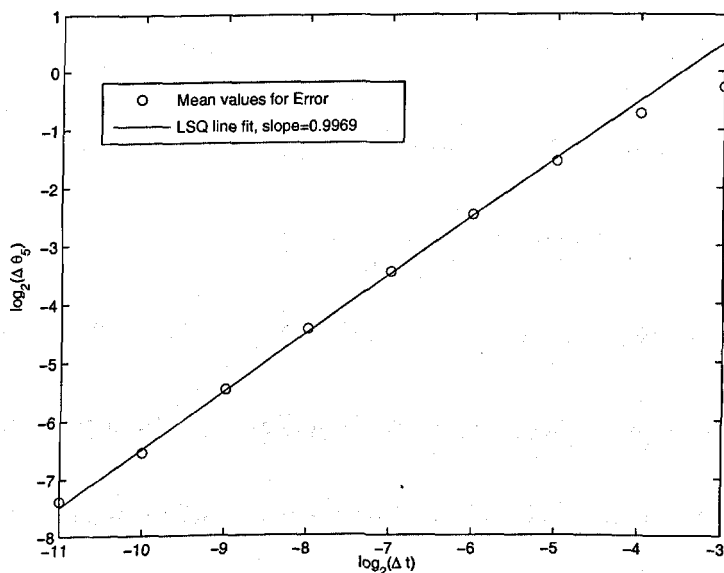


Figure 3.5: θ_5 fit for polynomial potential

The drift parameters obtained from (3.20) are averaged over $N_s = 1000$ realisations for a time interval $[0, T_f]$. Plotting the deviation of the estimated drift parameters $\hat{\theta}_5$ and $\hat{\theta}_3$ from the true drift parameters θ_5 and θ_3 the plots in figures 3.5 and 3.6 are obtained. The convergence observed here is representative of the whole parameter set. Using a least squares fit, a straight line can be fitted to those datapoints corresponding to small Δt and the obtained slopes of 0.9969 and 0.9976 respectively corroborated the estimator being asymptotically consistent with a bias of $\mathcal{O}(\Delta t)$.

3.3.4 Drift Parameters: Trigonometric Potentials

As a first step towards fitting stochastic differential equations to molecular dynamics data, fitting the type of SDE given in (3.1) to the Langevin-trajectories for butane obtained in chapter 2 is attempted. In order to adapt the fitted SDE to the process at hand, a new set of basis functions based on a trigonometric potential is chosen:

$$\begin{aligned}
 V_i(x) &= \frac{1}{i} \cos^i(x) \\
 f_i(x) &= -\sin(x) \cos^{i-1}(x)
 \end{aligned}
 \tag{3.21}$$

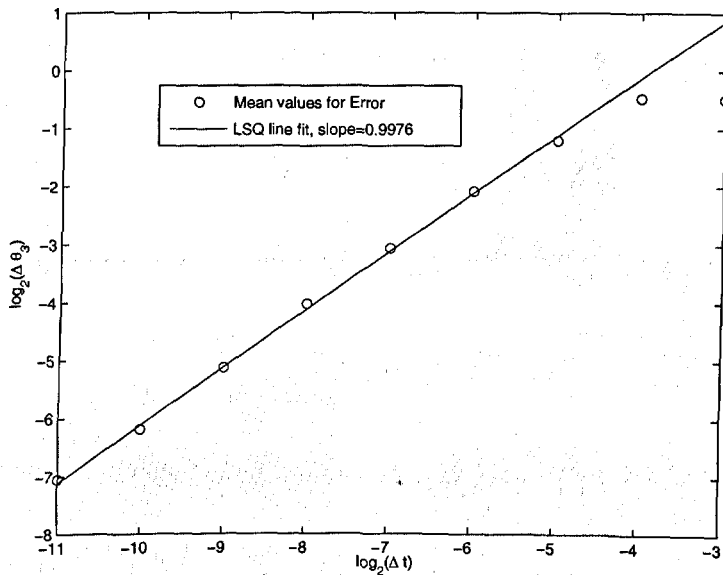


Figure 3.6: θ_3 fit for polynomial potential

The same fitting procedure based on the estimator (3.10) remains operational. To ensure the method of fitting is sound, we first consider parameter estimation for paths generated from the SDE. Since one unfavourable case, in which the errors observed decay more quickly than expected from an Ito-Taylor expansion, is encountered, this is analysed in some detail. An explicit analytical error expression is made available and compared to numerically obtained errors.

In order to find a suitable model case for parameter estimation, it is decided to mimic metastabilities using the following choice of coefficients:

$$\begin{aligned} \theta_1 &= 1 & \theta_2 &= 2 & \theta_3 &= 0 & \theta_4 &= 0 & \theta_5 &= 0 \\ \sigma &= 0.6 & N_s &= 100 & T_f &= 1.28 \cdot 10^5 \end{aligned}$$

This gives rise to the potential, typical sample path and histogram of figure 3.7.

For constant final time, the timestep is halved and the estimator is sampled $N_s = 100$ times for each time resolution, which yields the error plot given in figure 3.8. This plot corroborates the estimator being asymptotically consistent with errors of order $\mathcal{O}(\Delta t)$.

not quite sure what does this mean?

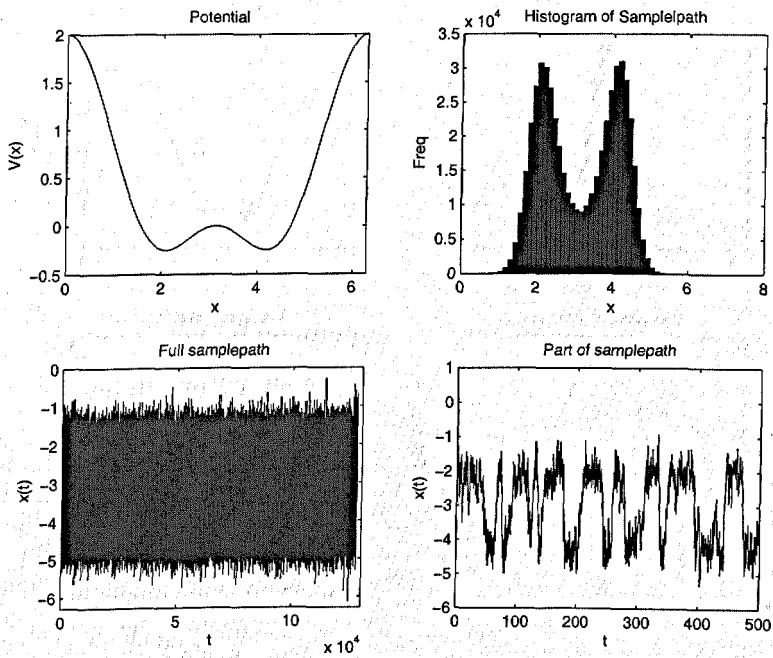


Figure 3.7: Trigonometric Potential and Sample Path

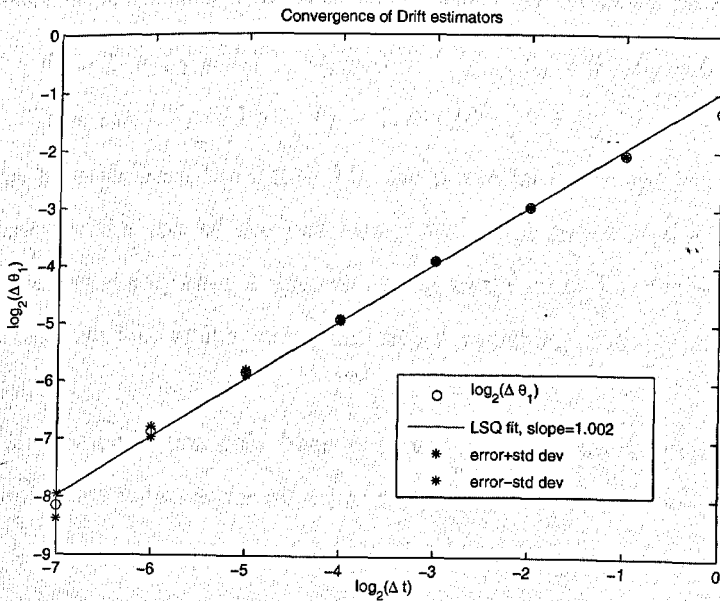


Figure 3.8: Convergence of $\hat{\theta}_1$

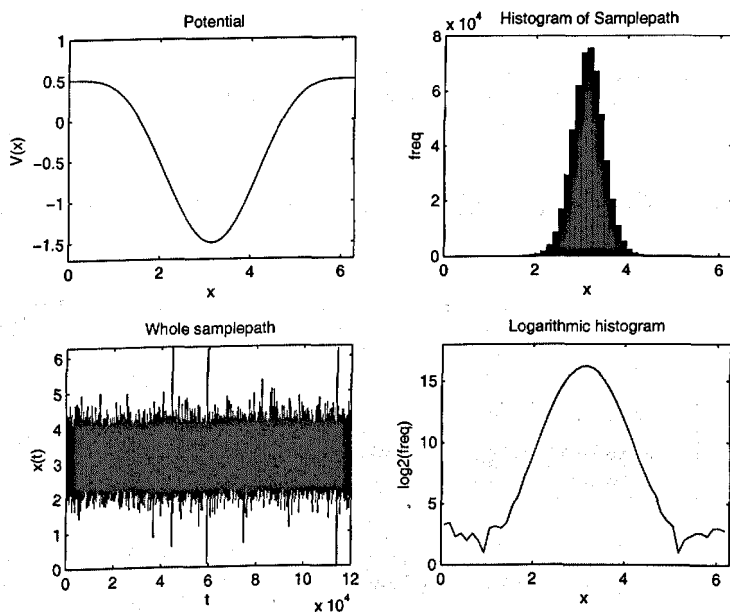


Figure 3.9: Trigonometric potential – unfavourable case

Unfortunately, there appear to be some combinations of parameters for which this convergence is not observed in practise. Consider the following selection:

$$\begin{aligned} \theta_1 = 1 \quad \theta_2 = -1 \quad \theta_3 = 0 \quad \theta_4 = 0 \quad \theta_5 = 0 \\ \sigma = 1.3 \quad N_s = 100 \quad T_f = 1.28 \cdot 10^5 \quad k = 2 \end{aligned}$$

As can be seen from the figure 3.9, the potential landscape is reasonably well-sampled except for the top of the potential. Also, the errors in the drift parameter estimators decreases – only that it does so at too fast a rate! The fastest rate achieved seems to increase with k , which might indicate a vanishing of lower order error terms due to symmetries.

This phenomenon has only been observed with the above or close by combinations of parameters whereas for all other tests, no such super-convergence has been found. While behaviour towards small values of Δt in 3.10 is slightly different for different random number generators, it should be stressed that the initial superconvergence is observed with three different random number generators, not including the built-in `rand()` which was ruled out previously.

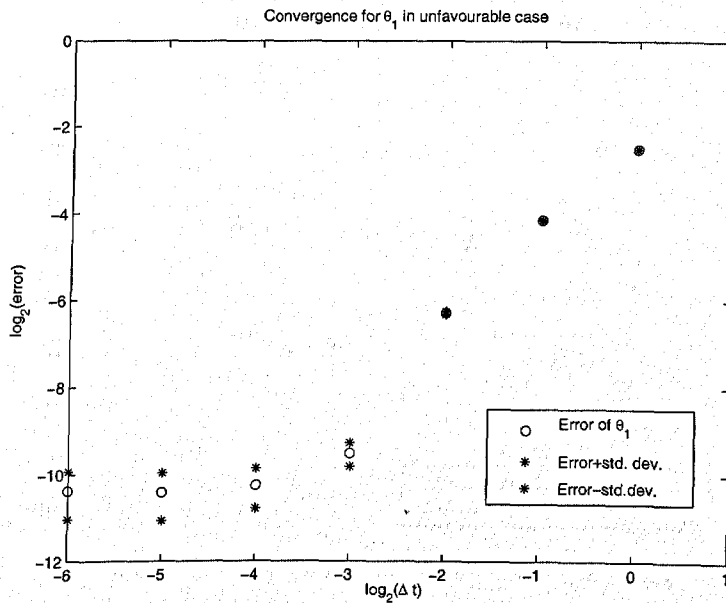


Figure 3.10: Low $\hat{\theta}_1$ Errors in unfavourable case

For the case $k = 2$ it is feasible to write out the actual statistics of the generated sample path and understand the first order error incurred by the Euler statistical model in detail. The analysis starts out by writing out the statistics for the $k = 2$ subsampled Euler-Maruyama generator:

$$x_{n+1} = x_n + \Delta t \sum_i \theta_i f_i(x_n) + \sigma \sqrt{\Delta t} \xi$$

$$+ \frac{1}{2} \Delta t \sum_i \theta_i f_i'(x_n) \cdot \left(\frac{1}{2} \Delta t \sum_j \theta_j f_j(x_n) + \sigma \sqrt{\frac{\Delta t}{2}} \xi^{(1)} \right) + \text{higher order terms.}$$

where ξ and $\xi^{(1)}$ are standard normal random variables (not independent). Using this representation those first order components of the error of the estimator

$$\Delta \theta = \hat{\theta} - \theta \quad (3.22)$$

can be expressed as a solution of the following linear system:

$$\mathbb{E}(M \Delta \theta)_j = \frac{1}{4} \Delta t \sum_{k,l} \sum_{n=0}^{N-1} \Delta t f_j(x_n) f_k(x_n) f_l'(x_n) + \text{h.o.t.} \quad (3.23)$$

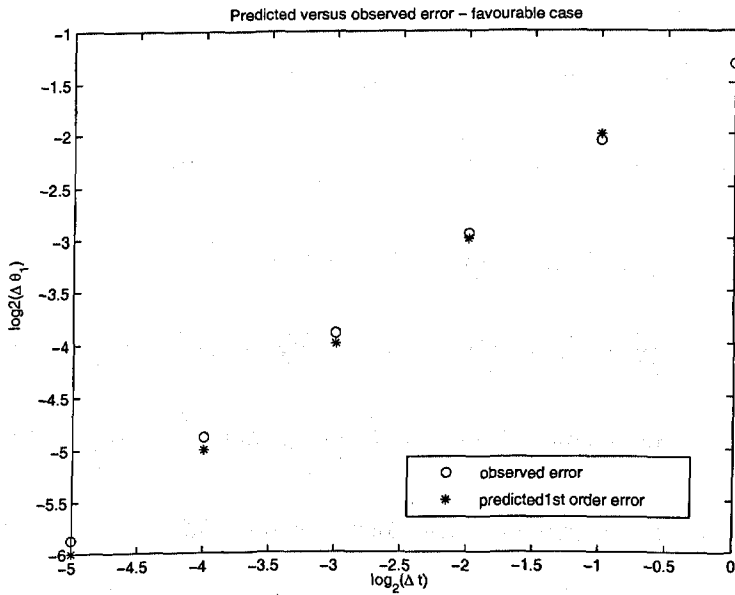


Figure 3.11: Predicted 1st Order Errors – favourable case

Here, the index $(M\Delta\theta)_j$ stands for the j -th entry of the vector $M\Delta\theta$ and the matrix M is as given in (3.19).

Since an ergodic theorem of the type

$$\lim_{\substack{N\Delta t \rightarrow \infty \\ \Delta t \rightarrow 0}} \frac{1}{N\Delta t} M_{i,j} = \int_0^{2\pi} f_j(x) f_i(x) d\mu(x) \quad (3.24)$$

with the invariant measure $\mu(\cdot)$ given by its density

$$\frac{d\mu}{dx} = \frac{1}{C(\theta, \sigma)} \exp\left(\frac{-2 \sum_i \theta_i V_i(x)}{\sigma^2}\right) \quad (3.25)$$

is expected to hold, it should be possible to write down integral expressions for $\Delta\theta$.

Since the basis functions f_i chosen here lead to a Fourier decomposition of the measure, assuming (3.24) one can show that e.g. the highest order coefficient, $\Delta\theta_4$ and also $\Delta\theta_1$ for the above problematic case must be zero.

Also, the above expression for the first order error correction has been implemented and found to agree well with the observed mean error in the case $\theta =$

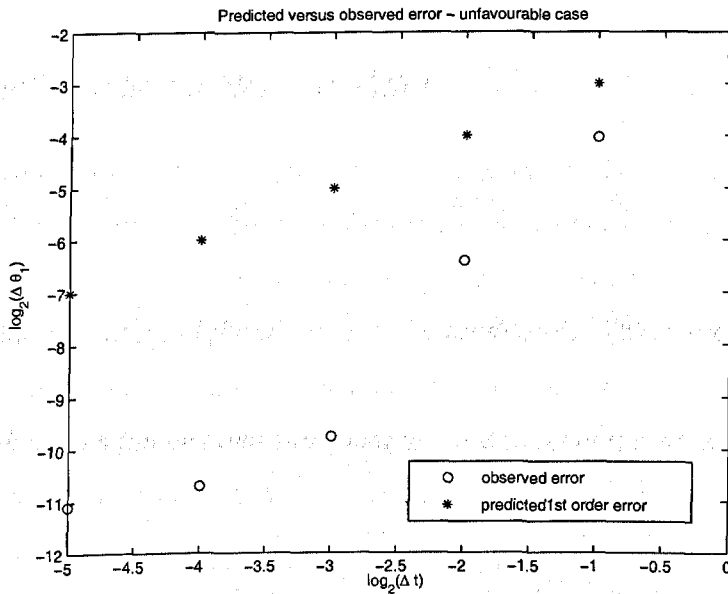


Figure 3.12: Predicted 1st Order Errors – unfavourable case

$[1, 2, 0, 0, 0]^T$ presented above. In the unfavourable case $\theta = [1, -1, 0, 0, 0]^T$, however, the predicted first order correction is well above the observed errors.

While it is conceivable that the error analysis only holds for much smaller Δt as it is only an asymptotic analysis after all, this argument is not very convincing especially in the presence of well-fitted analytical error expressions in the favourable case. Also, finite final time T together with a non-random starting point may introduce an additional bias which could spoil convergence of the parameter estimates to the true value. A complete resolution of this problem is left pending in order to enable the study of more pertinent problems.

3.3.5 Decay of Variance

The variances for the above estimators of the drift parameters are numerically observed to decay like $\mathcal{O}\left(\frac{1}{T_f}\right)$ which is expected from analogy with the law of large numbers.

Proving such a statement cannot be accomplished by mere Ito-Taylor expansions - some theorem of ergodic type must be used. There are some results in [50] to this

effect which are applicable to the trigonometric potential case.

3.4 Application to Butane data

Using the data shown in figure 2.15 it is possible to test fitting first order SDEs with the above trigonometric potential to molecular dynamics data – albeit in an extremely simple molecule. To get a first impression of the (in)consistency of such a model, fitting at different inter-sample spacings Δt is considered. This corresponds to only using every k th sample from the molecular dynamics simulation. The aim is to establish whether SDE models can be fitted convincingly over a range of timescales and one of the main problems for short timescales is the fact that the fitted paths have zero quadratic variation.

To start the study, consider the fit performed in figure 3.13 using the estimator (3.20) for the drift parameters and quadratic variation for the diffusion parameter.

It is clear that completely different potential expressions will be obtained for different sampling periods Δt . This is again shown in figure 3.14.

In particular, $\hat{\sigma} \rightarrow 0$ is observed as $\Delta t \rightarrow 0$. Of course, this is due to the smooth paths generated by the Hamiltonian dynamical system – the quadratic variation of those paths is zero!

In view of the fact that even if the data originates from a process of the kind (3.16) a small Δt is needed to control the bias this observation calls the fitting into question. For small timescales, the problems are inherent in the path to be fitted, whereas for long timescales, the fitting procedure is not sound. Either, different estimators are needed, or the process just cannot be fitted to the data. Even if the fit is successful with nearly constant drift parameters for some range of sampling rates Δt , weak convergence as in the situation of a distinguished particle in a heat bath is ruled out.

One way of fitting first order SDEs to molecular dynamics data at longer timescales might be to introduce imputed points between sampled datapoints which would then have to be sampled from, possibly alternating with samples from drift and diffusion parameters. *In extremis*, one could consider fitting a finite state space Markov chain at

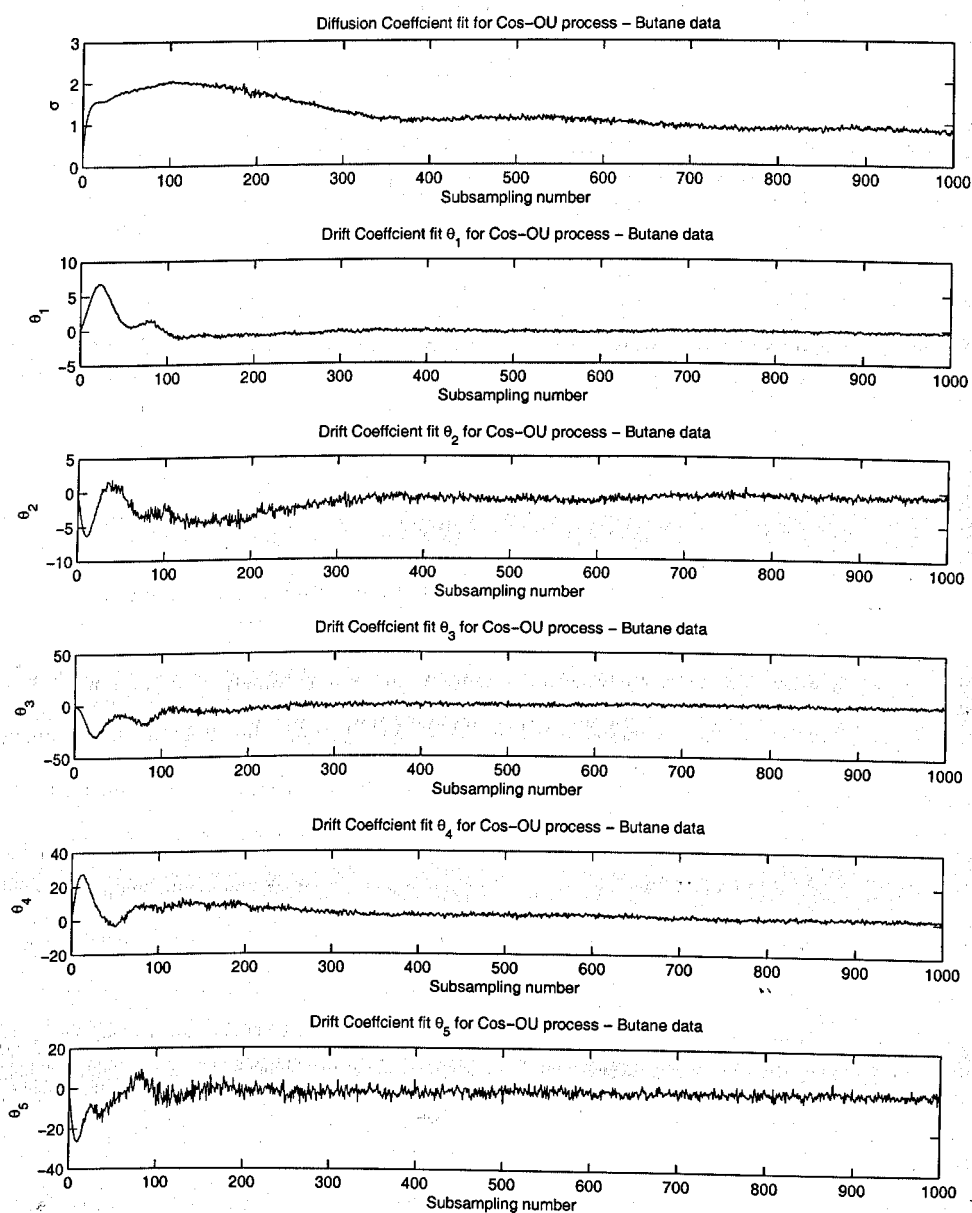


Figure 3.13: Fitting Trigonometric Potentials to Butane – Coefficients

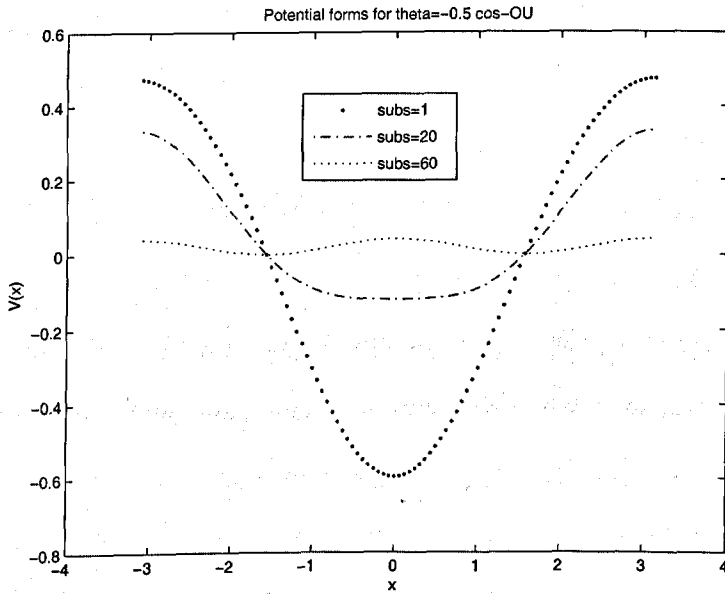


Figure 3.14: Different Potentials at Different k

very long time intervals, possibly corresponding to different conformational substates as advocated by Schütte et al. (see [35]). Rather than following this route, the investigation turns towards the small timescale structure of the data dealing with hypoelliptic diffusion processes in the next chapter.

3.5 The Two scale Potential

Another interesting example which has promise to be analytically tractable is the two-scale potential:

$$dx = -\frac{\partial}{\partial x} V(x, \varepsilon) dt + \sigma dB$$

where the potential V is such that it has a fast component which is averaged out:

$$V(x, \varepsilon) = \bar{V}(x) + \sin\left(\frac{x}{\varepsilon}\right)$$

Omitting some technical conditions on V , in this situation, it is known analytically that the solutions converge in some weak sense to the solutions of

$$dx = -\frac{\partial}{\partial x} \bar{V}(x) + \Sigma dB$$

where $\Sigma < \sigma$ seems intuitively reasonable.

Studying this process might provide some insight into how the estimators deal with the strong high frequency components visible in the data in figure 2.15. This study has been carried out for simple maximum likelihood estimators by Pavliotis and Stuart in [26], where subsampling was found necessary to avoid $\mathcal{O}(1)$ errors in the estimated parameters.

3.6 Conclusions

Standard results concerning maximum likelihood estimators for drift parameters and method of moment estimators for diffusivity have been summarised for a particular class of 1D SDEs. The approximating estimators have been implemented and tested on a variety of cases, in particular for trigonometric force expressions. These tested routines were then applied to molecular dynamics data for the dihedral angle in butane under Langevin dynamics and two issues of practical relevance were highlighted. Firstly, the fitted potential depends greatly on the timescale on which the fit is performed so that, even allowing for some error due to the finite Δt bias of the estimators used, no region of acceptable fit could be identified. Secondly, fitting for very short timescales is limited by the quadratic variation of the trajectory from molecular dynamics being zero. In this context, hypoelliptic diffusions might constitute an interesting class of processes to fit to this data and they are the subject of the next chapter.

Chapter 4

Second Order Stochastic Differential Equations

Wenn Sie eine Theorie haben, und Sie können den harmonischen Oszillator nicht rechnen, dann vergessen Sie sie.

Prof.Dr.H.D.Duebner, Clausthal

4.1 Overview

This chapter is a slightly enlarged version of [61] treating parameter estimation for partially observed hypoelliptic diffusion processes. By partial observation we mean observation of some components of the multidimensional process at discrete times. Since exact likelihoods for the transition densities are typically not known, approximations are used that are expected to work well in the limit of small inter-sample times Δt and large total observation times $N\Delta t$. Hypoellipticity together with partial observation leads to ill-conditioning requiring a judicious combination of approximate likelihoods for the various parameters to be estimated. We combine these in a deterministic scan Gibbs sampler alternating between missing data in the unobserved solution components, and parameters. Numerical experiments display asymptotic consistency of the method when applied to simulated data. The chapter concludes with application of the Gibbs sampler to molecular dynamics data generated as described in chapter 2.

4.2 Introduction

In many application areas it is of interest to model some components of a large deterministic system by a low dimensional stochastic model. In some of these applications, insight from the deterministic problem itself forces structure on the form of the stochastic model, and this structure must be reflected in parameter estimation. In this chapter, we study the fitting of stochastic differential equations (SDEs) to discrete time series data in situations where the model is a hypoelliptic diffusion process,¹ and also where observations are only made of variables that are not directly forced by white noise. Such a structure arises naturally in a number of applications.

One application is the modeling of macro-molecular systems [32] and [34]. In its basic form the molecule is described by a large Hamiltonian system of ordinary differential equations (ODEs). If the molecule spends most of its time in a small number of macroscopic configurations then it may be appropriate to model the dynamics within, and in some cases between, these states by a hypoelliptic diffusion. While this phrasing of the question is relatively recent, under the name of the "Kramers problem" it dates back to [41] with a brief summary in section 5.3.6a of [9]. As observed in the last chapter, the trajectories generated by the Hamiltonian mechanics model of molecular dynamics are smooth which compromises the fitting of first order stochastic differential equations. Moving to second or higher order hypoelliptic SDEs might enlarge the range of timescales useable for fitting. Furthermore, inertial effects are present even in large molecules in solution - otherwise infrared spectroscopic observations would be entirely meaningless, so it might be desirable to model them. Another application, audio signal analysis, is referred to in [30] where a continuous time ARMA model is used.

We consider SDE models of the form

$$\begin{cases} dx = \Theta A(x)dt + CdB \\ x(0) = x_0 \end{cases} \quad (4.1)$$

where B is an m -dimensional Wiener process and x a k -dimensional continuous process with $k > m$. $A : \mathbb{R}^k \rightarrow \mathbb{R}^l$ is a set of (possibly non-linear) globally Lipschitz force

¹Meaning that the covariance matrix of the noise is degenerate, but the probability densities are smooth.

functions. The parameters which we estimate are the last m rows of the drift matrix, $\Theta \in \mathbb{R}^{k \times l}$, and the diffusivity matrix C which we assume to be of the form

$$C = \begin{bmatrix} 0 \\ \Gamma \end{bmatrix} \in \mathbb{R}^{k \times m}$$

where $\Gamma \in \mathbb{R}^{m \times m}$ is nonsingular.

It is known that under the above hypotheses on A and C , a unique L^2 -integrable solution $x(\cdot)$ exists almost-surely for all times $t \in \mathbb{R}^+$, see e.g. Theorem 5.2.1 in [3]. We also assume that the process defined by (4.1) is hypoelliptic as defined in [48], i.e. it satisfies Hörmander's hypothesis as given in section V.38 of [42]. Intuitively, this corresponds to the noise being spread into all components of the system (4.1) via the drift.

The structure of C implies that the noise acts directly only on a subset of the variables which we refer to as *rough*. It may then be transmitted, through the coupling in the drift, to the remaining parts of the system which we refer to as *smooth*². To distinguish between rough and smooth variables, we introduce the notation $x(t)^T = (u(t)^T, v(t)^T)$ where $u(t) \in \mathbb{R}^{k-m}$ is smooth and $v(t) \in \mathbb{R}^m$ is rough. It is helpful to define linear functions $P : \mathbb{R}^k \rightarrow \mathbb{R}^{k-m}$ by $Px = u$ and $Q : \mathbb{R}^k \rightarrow \mathbb{R}^m$ by $Qx = v$.

We denote the sample path at $N + 1$ equally spaced points in time by $\{x_n = x(n\Delta t)\}_{n=0}^N$, and we write $x_n^T = (u_n^T, v_n^T)$ to separate the rough and smooth components. Also, for any sequence (z_1, \dots, z_N) , $N \in \mathbb{N}$ we write $\Delta z_n = z_{n+1} - z_n$ to denote forward differences. We are mainly interested in cases where only the smooth component, u , is observed and our focus is on parameter estimation for all of Γ and for entries of those rows of Θ corresponding to the rough path, on the assumption that $\{u_n\}_{n=0}^N$ are samples from a true solution of (4.1); such a parameter estimation problem arises naturally in many applications and an example is given in section 4.8. It is natural to consider $N\Delta t = T \gg 1$ and $\Delta t \ll 1$. It is important to realize that, for continuous time observations, the diffusion coefficient Γ can be found from the quadratic variation of a single path on $[0, T]$, any $T > 0$, see e.g. Theorem 2.8.6 in [13]. For Θ , however, the estimates are strongly consistent only as $T \rightarrow \infty$. These two facts will be reflected

²We do not mean C^∞ here, but they are at least C^1 .

in the parameter estimation for discrete time observations.

The sequence $\{x_n\}_{n=0}^N$ defined above is generated by a Markov chain. By expanding the random map $x_n \mapsto x_{n+1}$ in powers of Δt , and retaining the leading order contributions to the mean and to the variance in each component of the equation, one obtains

$$x_{n+1} \approx x_n + \Delta t \Theta A(x_n) + \sqrt{\Delta t} R(\Delta t; \Theta) \xi_n \quad (4.2)$$

where $x_n \in \mathbb{R}^k$, $\xi_n \in \mathbb{R}^k$ is distributed as $\mathcal{N}(0, I)$ and $R(\Delta t; \Theta) \in \mathbb{R}^{k \times k}$. Because of the hypoellipticity, $R(\Delta t; \Theta)$ is invertible, but the zeros in C mean that it is highly ill-conditioned for $0 < \Delta t \ll 1$. In fact we have:

$$R(0; \Theta) = \begin{bmatrix} 0 & 0 \\ 0 & \Gamma \end{bmatrix}. \quad (4.3)$$

We refer to expressions of the form (4.2) as statistical models and we will use them to approximate the exact likelihood, $\mathcal{L}(u, v | \Theta, \Gamma \Gamma^T)$, of the path u, v given parameter values Θ and $\Gamma \Gamma^T$.

Given prior distributions for the parameters, $p_0(\Theta, \Gamma \Gamma^T)$, the posterior likelihood can be constructed as follows:

$$\begin{aligned} \mathcal{L}(v, \Theta, \Gamma \Gamma^T) &= \frac{\mathcal{L}(v, \Theta, \Gamma \Gamma^T, u)}{\mathcal{L}(u)} \\ &= \mathcal{L}(u, v | \Theta, \Gamma \Gamma^T) \frac{p_0(\Theta, \Gamma \Gamma^T)}{\mathcal{L}(u)} \end{aligned} \quad (4.4)$$

In principle, this can be used as the basis for Bayesian sampling of $(\Theta, \Gamma \Gamma^T)$, viewing v as missing data. However, the exact likelihood of the path is typically unavailable. In this chapter we will combine judicious approximations of this likelihood to solve the sampling problem. The approximations that we use, \mathcal{L}_E and \mathcal{L}_{IT} , are found from (4.2), in the case of \mathcal{L}_E by replacing $R(\Delta t; \Theta)$ with $R(0; \Theta)$ given by (4.3). Thus \mathcal{L}_E is found from an Euler-Maruyama approximation of (4.1). The approximate likelihood \mathcal{L}_{IT} arises from retaining further terms in the Itô-Taylor expansion to ensure that noise is propagated into each component of the map (4.2).

The questions we address in this chapter are:

1. How does the ill-conditioning of the Markov chain $x_n \mapsto x_{n+1}$ affect parameter estimation for $\Gamma \Gamma^T$ and for the last m rows of Θ in the regime $\Delta t \ll 1$, $N \Delta t = T \gg 1$?

2. In many applications, it is natural that only the smooth data $\{u_n\}_{n=0}^N$ is observed, and not the rough data $\{v_n\}_{n=0}^N$. What effect does the absence of observations of the rough data have on the estimation for $\Delta t \ll 1$ and $N\Delta t = T \gg 1$?
3. The exact likelihood is usually not available; what approximations of the likelihood should be used, in view of the ill-conditioning?
4. How should the answers to these questions be combined to produce an effective method to sample the distribution of parameters Θ, Γ^T and the missing data $\{v_n\}_{n=0}^N$?

To tackle these issues, we use a combination of analysis and numerical simulation, based on three model problems which are conceived to highlight issues central to the questions above. We will use analysis to explain why some seemingly reasonable methods fail, and simulation will be used both to extend the validity of the analysis and to illustrate good behavior of other methods.

For the numerical simulations, we will use either exact discrete time samples of (4.1) in simple linear cases, or trajectories obtained by Euler-Maruyama simulation of the SDE on a temporal grid with a spacing considerably finer than the observation time interval Δt .

At this point, we introduce some notation to simplify the presentation. Firstly, given an invertible matrix $R \in \mathbb{R}^{n \times n}$ we introduce a new norm using the Euclidean norm on \mathbb{R}^n by setting $\|x\|_R = \|R^{-1}x\|_2$ for vectors $x \in \mathbb{R}^n$. Also, we will occasionally refer to a likelihood $\mathcal{L}(B)$ as a function of some parameters B not mentioning the complementary parameter set C . This is understood to refer to the conditional likelihood $\mathcal{L}(B|C)$ whenever the parameter set C is clear from the context.

In section 2 we will introduce our three model problems and in section 3 we study the performance of \mathcal{L}_E to estimate the diffusion coefficient. Observing and analysing its failure in the case with partial observation leads to the improved statistical model yielding \mathcal{L}_{IT} which eliminates these problems; we introduce this in section 4. In section 5 we show that \mathcal{L}_{IT} is inappropriate for drift estimation, but that \mathcal{L}_E is effective in this context. In section 6, the individual estimators will be combined into a Gibbs sampler to

solve the overall estimation problem with asymptotically consistent performance being demonstrated numerically. Section 7 contains a simple application to molecular dynamics and section 8 provides concluding discussion.

4.2.1 Literature review

The primary novelty of our work is that it concerns hypoelliptic diffusions where only smooth components are observed. We set our work in context. First, we review parameter estimation for (4.1) in continuous time. We assume that the observation is compatible with (4.1) in that, if the observed path is $x(t)^T = (u(t)^T, v(t)^T)$, then

$$\dot{u} = P\Theta A(x), \quad u(0) = Px(0); \quad (4.5)$$

furthermore, if only $u(t)$ is observed, then we assume that (4.5) determines $v(t)$ uniquely. (In situations where compatibility fails it is necessary to add observational noise to the solution of (4.5) and to estimate it.)

Once v is determined uniquely we have

$$dv = Q\Theta A(x) + \Gamma dB, \quad v(0) = Qx(0). \quad (4.6)$$

The covariance matrix $\Gamma\Gamma^T$ can be estimated by noting that

$$\frac{1}{T} \sum_{n=0}^{N-1} (v_{n+1} - v_n)(v_{n+1} - v_n)^T \rightarrow \Gamma\Gamma^T \quad \text{as } N \rightarrow \infty \quad (4.7)$$

with $T = N\Delta t$ fixed [13].

The Girsanov formula shows that the path space likelihood for (4.6) is proportional to

$$\exp \left(\int_0^T \Gamma^{-1} Q \Theta A(x(s)) \Gamma^{-1} dv(s) - \frac{1}{2} \int_0^T \|\Gamma^{-1} Q \Theta A(x(s))\|^2 ds \right).$$

This can be used as the basis for various estimation procedures, one of them being the maximum likelihood estimator for the lower rows of Θ which is found by maximizing

$$\hat{\Theta} = \operatorname{argmax}_{\Theta} \left(\int_0^T \Gamma^{-1} Q \Theta A(x(s)) \Gamma^{-1} dv(s) - \frac{1}{2} \int_0^T \|\Gamma^{-1} Q \Theta A(x(s))\|^2 ds \right) \quad (4.8)$$

over Θ . Such estimators are consistent as $T \rightarrow \infty$. In the linear case, where A is just the identity, the maximum likelihood estimate for the whole of Θ is given by

$$\hat{\Theta} = \left[\int_0^T dx x^T \right] \left[\int_0^T x x^T dt \right]^{-1}. \quad (4.9)$$

This is proved to be consistent as $T \rightarrow \infty$ in [4]. Note, then, that diffusion parameters can be estimated from arbitrarily short pieces of trajectory, whereas drift parameters require long time intervals. A discussion of continuous time parameter estimation for linear hypoelliptic diffusions with multiplicative noise is given in [39].

In practice, observations are typically made in discrete time. There is substantial literature on parameter estimation in this context, much of it concerned with estimation of φ in problems of the form

$$dv = a(v, \varphi)dt + \Gamma \dot{w}, \quad v(0) = v_0, \quad (4.10)$$

where $\Gamma \Gamma^T$ is everywhere invertible. In some cases, a is allowed to depend on the entire path $\{v(s)\}_{s \in [0, t]}$ and then the hypoelliptic problem (4.6) is a special case. We now discuss the literature available when only discrete time observations of v , the rough variable, are given. Note that, for most of this chapter, we assume that the v -data is hidden and only u in (4.1) is observed. Thus although u can be eliminated from (4.1), and an equation written for v in the form (4.10) with a depending on the entire path of v on $[0, t]$, the existing literature on discrete time observations of (4.10) does not apply to the case we consider here, where v is not observed. Nonetheless we overview what is known.

One approach is to form continuous time estimators, using the generalization of (4.8) to (4.10). If φ appears linearly and only in a , not γ , then the continuous time estimator can be calculated from Riemann and stochastic integrals of $v(t)$. These continuous time estimators can be approximated by quadrature, assuming the time increment between observations, Δt , is small, and estimates of $\hat{\varphi}$ obtained in this manner, see [40] for details. An alternative, when Δt is small, is to approximate the likelihood of the discrete time Markov chain generated by sampling (4.10) at rate Δt . This approach is considered in [55, 29, 11, 24] with several of these papers studying the Euler approximation, generating a Gaussian likelihood, as we do in this chapter. Theorems about

convergence of parameter estimates typically consider the limit $\Delta t \rightarrow 0$ with $N\Delta t \rightarrow \infty$ [24]. Alternatively one may consider $\Delta t \rightarrow 0$ with $N\Delta t = T \gg 1$ and estimate the bias due to finite T .

When the time increment between observations, Δt , is not small then $O(1)$ errors can enter parameter estimates unless the discrete time likelihood is carefully approximated. One way to do this is by fine Monte Carlo simulation between observation points, see [49]. A different approach, leading to closed formulas and using Hermite polynomials, may be found in [1]. In [11] functionals of the Brownian bridge are used to build up the approximation; in [53] related ideas are used in a Bayesian approach to parameter estimation for discretely observed diffusions. Recent work of Beskos et al uses exact sampling of a diffusion process to address this issue, see [16]. Another approach is taken by Crommelin and Vanden-Eijnden in [7], [8] in which the transition probability matrix is approximated from the data, and then a generator is found to fit the spectrum of that matrix as closely as possible. The norm used to facilitate fitting is such that quadratic programming techniques can be used to speed up computation. A review of estimation for discretely observed diffusion processes, and a discussion of martingale estimating functions, can be found in [2].

4.3 Model Problems

To study the performance of parameter estimators, we have selected a sequence of three Model Problems ranging from simple linear stochastic growth through a linear oscillator subject to noise and damping to a nonlinear oscillator of similar form. All these problems are hypoelliptic diffusions and we will present them in detail in the next three subsections. Their general form is given as the second order Langevin equation

$$\begin{cases} dq = p dt, \\ dp = (-\gamma p + f(q)) dt + \sigma dB \end{cases} \quad (4.11)$$

where f is some (possibly nonlinear) force-function and the variables q and p are scalar.

4.3.1 Model Problem I: Stochastic Growth

Here, $x = (q, r)^T$ satisfies

$$\begin{cases} dq = rdt \\ dr = \sigma dB. \end{cases} \quad (4.12)$$

The process has one parameter, the diffusion parameter σ , that describes the size of the fluctuations. In the setting of (4.1) we have

$$A(x) = x, \quad \Theta = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}$$

and $u = q, v = r$. The process is Gaussian with mean and covariance

$$\mu(t) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} q_0 \\ r_0 \end{bmatrix} \quad \text{and} \quad \Sigma(t) = \sigma^2 \begin{bmatrix} t^3/3 & t^2/2 \\ t^2/2 & t \end{bmatrix}.$$

The exact discrete samples may be written as

$$\begin{cases} q_{n+1} = q_n + r_n \Delta t + \sigma \frac{(\Delta t)^{3/2}}{\sqrt{12}} \zeta_n^{(1)} + \sigma \frac{(\Delta t)^{3/2}}{2} \zeta_n^{(2)}, \\ r_{n+1} = r_n + \sigma \sqrt{\Delta t} \zeta_n^{(2)}, \end{cases} \quad (4.13)$$

with $\zeta_0 \sim \mathcal{N}(0, I)$ and $\{\zeta_n\}_{n=0}^N$ being i.i.d.

4.3.2 Model Problem II: Harmonic Oscillator

As our second model problem we consider a damped harmonic oscillator driven by a white noise forcing where $x = (q, p)^T$:

$$\begin{cases} dq = pdt \\ dp = -Dqdt - \gamma pdt + \sigma dB. \end{cases} \quad (4.14)$$

This model is obtained from the general SDE (4.1) for the choice

$$A(x) = x, \quad \Theta = \begin{bmatrix} 0 & 1 \\ -D & -\gamma \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}$$

and $u = q, v = p$. The process is Gaussian and the mean and covariance of the solution can be explicitly calculated.

4.3.3 Model Problem III: Oscillator with trigonometric potential

In the third model problem, $x = (q, p)^T$ describes the dynamics of a particle moving in a potential which is a superposition of trigonometric functions and in contact with a heat bath obeying the fluctuation-dissipation relation, see [43]. This potential is sometimes used in molecular dynamics in connection with the dynamics of dihedral angles – see section 4.8. The model is

$$\begin{cases} dq = p dt, \\ dp = (-\gamma p - \sum_{j=1}^c D_j \sin(q) \cos^{j-1}(q)) dt + \sigma dB. \end{cases} \quad (4.15)$$

This equation has parameters γ , D_i , $i = 1, \dots, c$ and σ . It can be obtained from the general SDE (4.1) for the choice

$$A \left(\begin{bmatrix} q \\ p \end{bmatrix} \right) = \begin{bmatrix} \sin(q) \\ \sin(q) \cos(q) \\ \vdots \\ \sin(q) \cos^{c-1}(q) \\ p \end{bmatrix}, \quad \Theta = \begin{bmatrix} 0 & \dots & 0 & 1 \\ -D_1 & \dots & -D_c & -\gamma \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ \sigma \end{bmatrix}.$$

and $u = q$, $v = p$. No explicit closed-form expression for the solution of the SDE is known in this case; the process is not Gaussian.

4.4 Euler Statistical Model

In this section, the Euler-Maruyama approximation to (4.1) is used to generate a statistical model and associated likelihood. Using this likelihood to estimate the diffusivity works whenever observations of both the smooth and the rough components are available. However, it yields $\mathcal{O}(1)$ errors in the partially observed case; this is demonstrated analytically for Model Problem I and the results are extended by means of numerical experiments.

4.4.1 Statistical Model

If the force function $A(\cdot)$ is nonlinear, closed-form expressions for the likelihood are in general unavailable. To overcome this obstacle, one can use a discrete time statistical

model. The Euler model is commonly used and we apply it to a simple linear model problem to highlight its deficiencies in the case of partially observed data from hypoelliptic diffusions.

The Euler-Maruyama approximation of the SDE (4.1) is

$$X_{n+1} = X_n + \Delta t \Theta A(X_n) + \sqrt{\Delta t} C \xi_n \quad (4.16)$$

where $\xi_n \sim \mathcal{N}(0, I)$ is an i.i.d. sequence of k -dimensional vectors with standard normal distribution. This corresponds to (4.2) with $R(\Delta t; \Theta)$ replaced by $R(0; \Theta)$ from (4.3).

Thus we obtain

$$\left\{ \begin{array}{l} U_{n+1} = U_n + \Delta t P \Theta A(X_n) \\ V_{n+1} = V_n + \Delta t Q \Theta A(X_n) + \sqrt{\Delta t} \Gamma \xi_n \end{array} \right\} \quad (4.17)$$

where now each element of the i.i.d. sequence ξ_n is distributed as $\mathcal{N}(0, I)$ in \mathbb{R}^m . This model gives rise to the following likelihood:

$$\mathcal{L}_{ND}(U, V | \Theta, \Gamma \Gamma^T) = \prod_{n=0}^{N-1} \frac{\exp(-\frac{1}{2} \|\Delta V_n - \Delta t Q \Theta A(X_n)\|_{\Gamma}^2)}{\sqrt{2\pi |\Gamma \Gamma^T|}} \delta\left(\frac{U_{n+1} - U_n}{\Delta t} - P \Theta A(X_n)\right) \quad (4.18)$$

The Dirac mass insists that the data is compatible with the statistical model, i.e. the V path must be given by numerical differentiation (ND) of the U path. To estimate parameters we will use the following expression:

$$\mathcal{L}_E(U, V | \Theta, \Gamma \Gamma^T) = \prod_{n=0}^{N-1} \frac{\exp(-\frac{1}{2} \|\Delta V_n - \Delta t Q \Theta A(X_n)\|_{\Gamma}^2)}{\sqrt{2\pi |\Gamma \Gamma^T|}}, \quad (4.19)$$

where we assume that $\{U_n\}, \{V_n\}$ are related through numerical differentiation when the Euler model is used to estimate missing components.

4.4.2 Model Problem I

The Euler statistical model for this model problem is

$$\left\{ \begin{array}{l} Q_{n+1} = Q_n + R_n \Delta t, \\ R_{n+1} = R_n + \sigma \sqrt{\Delta t} \xi_n. \end{array} \right. \quad (4.20)$$

Here, $\{\xi_n\}$ is an i.i.d. $\mathcal{N}(0, 1)$ sequence. The root cause of the phenomena that we discuss in this chapter is manifest in comparing (4.13) and (4.20). The difference is that

the $O((\Delta t)^{3/2})$ white noise contributions in the exact time series (4.13) do not appear in the equation for Q_n . We will see that this plays havoc with parameter estimation, even though the Euler method is pathwise convergent.

We assume that observations of the smooth component only, Q_n , are available. In this case the Euler method for estimation (4.20) gives the formula

$$R_n = \frac{Q_{n+1} - Q_n}{\Delta t} \quad (4.21)$$

for the missing data. In the following numerical experiment we generate exact data from (4.13) using the parameter value $\sigma = 1$. We substitute R_n given by (4.21) into (4.19) and find the maximum likelihood estimator for σ in the case of partial observation. In the case of complete observation we use the exact value for $\{R_n\}$, from (4.13), and again use a maximum likelihood estimator for σ from (4.19).

Using $N = 100$ timesteps for a final time of $T = 10$ with $\sigma = 1$ the histograms for the estimated diffusion coefficient presented in the middle column of Figure 4.4.2 are obtained. The top row contains histograms obtained in the case of complete observation where good agreement between the true σ and the estimates is observed. The bottom row contains the histograms obtained for partial observation using (4.21). The observed mean value of $\mathbb{E}\hat{\sigma} = 0.806$ indicates that the method yields biased estimates. Increasing the final time to $T = 100$ (see left column of graphs in Figure 4.4.2) or increasing the resolution to $\Delta t = 0.01$ do not remove this bias.

Thus we see that, in the case of partial observation, $\hat{\sigma}$ contains $O(1)$ errors which do not diminish with decreasing Δt and/or increasing $T = N\Delta t$.

4.4.3 Analysis of why the missing data method fails

Model Problem I can be used to illustrate why this method fails. We first argue that the method works without hidden data. The log-likelihood function given in (4.19) yields the following expression in the case of stochastic growth:

$$\log \mathcal{L}_E(\sigma) = -2N \log \sigma - \frac{1}{\sigma^2 \Delta t} \sum_{n=0}^{N-1} (\Delta r_n)^2$$

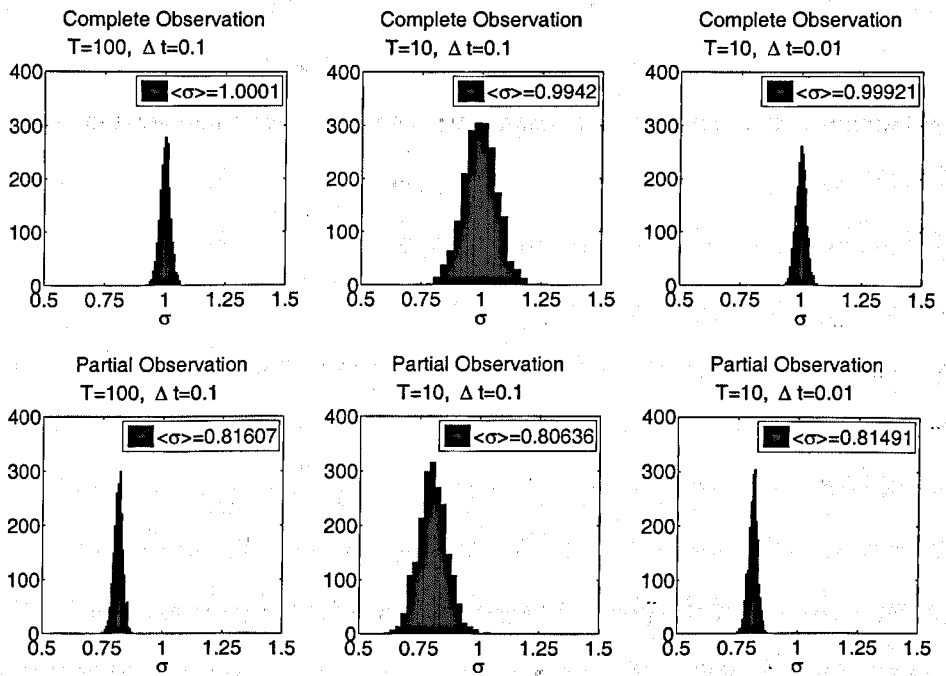


Figure 4.1: Estimates of σ using Euler Model for Model Problem I.
 Top row: fully observed process; bottom row: partially observed process.

where Δ is the forward difference operator. The maximum of the log-likelihood function gives the maximum likelihood estimate,

$$\hat{\sigma}^2 = \frac{1}{N\Delta t} \sum_{n=0}^{N-1} (\Delta r_n)^2. \quad (4.22)$$

In the case of complete data, (4.13) gives

$$\hat{\sigma}^2 = \frac{\sigma^2}{N} \sum_{n=0}^{N-1} (\zeta_n^{(2)})^2. \quad (4.23)$$

By the law of large numbers, $\hat{\sigma}^2 \rightarrow \sigma^2$ almost surely as $N \rightarrow \infty$. This shows that the method works when the complete data is observed.

Let us consider what happens when r is hidden. In this case, r_n is estimated by

$$\hat{r}_n = \frac{q_{n+1} - q_n}{\Delta t}.$$

But since q_n is generated by (4.13) we find that

$$\hat{r}_n = \frac{r_{n+1} + r_n}{2} + \sigma \frac{\sqrt{\Delta t}}{\sqrt{12}} \zeta_n^{(1)}$$

and

$$\begin{aligned} \Delta \hat{r}_n &= \frac{\Delta r_{n+1}}{2} + \frac{\Delta r_n}{2} + \sigma \frac{\sqrt{\Delta t}}{\sqrt{12}} (\zeta_{n+1}^{(1)} - \zeta_n^{(1)}) \\ &= \frac{\sigma \sqrt{\Delta t}}{2} \left(\zeta_{n+1}^{(2)} + \zeta_n^{(2)} + \frac{1}{\sqrt{3}} \zeta_{n+1}^{(1)} - \frac{1}{\sqrt{3}} \zeta_n^{(1)} \right) \end{aligned}$$

When $\Delta \hat{r}_n$ is inserted in (4.22) it follows that

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sigma^2}{4N} \sum_{n=0}^{N-1} \left(\zeta_{n+1}^{(2)} + \zeta_n^{(2)} + \frac{\zeta_{n+1}^{(1)} - \zeta_n^{(1)}}{\sqrt{3}} \right)^2 \\ &= \frac{\sigma^2}{4N} \left[\sum_{n=0}^{N-1} \left(\zeta_{n+1}^{(2)} + \frac{\zeta_{n+1}^{(1)}}{\sqrt{3}} \right)^2 + \sum_{n=0}^{N-1} \left(\zeta_n^{(2)} - \frac{\zeta_n^{(1)}}{\sqrt{3}} \right)^2 \right. \\ &\quad \left. + 2 \sum_{n=0}^{N-1} \left(\zeta_n^{(2)} - \frac{\zeta_n^{(1)}}{\sqrt{3}} \right) \left(\zeta_{n+1}^{(2)} + \frac{\zeta_{n+1}^{(1)}}{\sqrt{3}} \right) \right]. \end{aligned}$$

The random variables $\{\zeta_n\}_{n=0}^N$ are i.i.d with $\zeta_0 \sim N(0, I)$. So, by the law of large numbers, $\hat{\sigma}^2 \rightarrow \frac{2}{3}\sigma^2$ almost surely as $N \rightarrow \infty$. Furthermore, the limits hold in either of the cases where either $N\Delta t = T$ or Δt are fixed as $N \rightarrow \infty$. This means that

independently of what limit is considered, a seemingly reasonable estimation scheme based on Euler approximation results in $O(1)$ errors in the diffusion coefficient. ³

4.5 Improved statistical model

The failure of the Euler model to estimate paths having the correct quadratic variation is caused by not propagating the noise to the smooth component of the solution. A new statistical model is thus proposed which propagates the noise using what amounts to an Itô-Taylor expansion, retaining the leading order component of the noise in each row of the equation. The model is used to set up an estimator for the missing path using a Langevin sampler from path-space which is then simplified to a direct sampler in the Gaussian case. Numerical experiments indicate that the method yields the correct quadratic variation for the simulated missing path.

The model is motivated using our common framework the Model Problems I, II and III, namely (4.11). The improved statistical model is based on the observation that in the second row of an Itô-Taylor expansion of (4.11) the drift terms are of size $O(\Delta t)$ whereas the random forcing term is "typically" (in root mean square) of size $O(\sqrt{\Delta t})$. Thus, neglecting the contribution of the drift term in the second row on the first row leads to the following approximation of (4.11):

$$\begin{bmatrix} Q_{n+1} \\ P_{n+1} \end{bmatrix} = \begin{bmatrix} Q_n \\ P_n \end{bmatrix} + \Delta t \begin{bmatrix} P_n \\ f(Q_n) - \gamma P_n \end{bmatrix} + \sigma \begin{bmatrix} \int_0^{\Delta t} B(s) ds \\ B(\Delta t) \end{bmatrix}$$

The random vector on the right hand side is Gaussian, and can be expressed as a linear combination of two independent normally distributed Gaussian random variables. Computation of the variances and the correlation is straightforward leading to the following statistical model:

$$\begin{bmatrix} Q_{n+1} \\ P_{n+1} \end{bmatrix} = \begin{bmatrix} Q_n \\ P_n \end{bmatrix} + \Delta t \begin{bmatrix} P_n \\ f(Q_n) - \gamma P_n \end{bmatrix} + \sigma \sqrt{\Delta t} R \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \quad (4.24)$$

³There is similarity here with work of Gaines and Lyons [28] showing that adaptive methods for SDEs get the quadratic variation wrong if the adaptive strategy is not chosen carefully.

Here, ξ_1 and ξ_2 are normally distributed Gaussian random variables and R is given as

$$R = \begin{bmatrix} \frac{\Delta t}{\sqrt{12}} & \frac{\Delta t}{2} \\ 0 & 1 \end{bmatrix}$$

This is a specific instance of (4.2). It should be noted that this model is in agreement with the Ito-Taylor approximation up to error terms of order $\mathcal{O}(\Delta t^2)$ in the first row and $\mathcal{O}(\Delta t^{\frac{3}{2}})$ in the second row.

If complete observations are available, this model performs satisfactorily for the estimation of σ . This can be verified analytically for Model Problem I in the same fashion as in section 4.4.3. Numerically, this can be seen from the first row (referring to complete observation) of Figure 4.2 for Model Problem I and from the first row of Figure 4.3 for Model Problem II. In both cases the true value is given by $\sigma = 1$. See subsection 4.2 for a full discussion of these numerical experiments.

If only partial observations are available, however, a means of reconstructing the hidden component of the path must be procured. A standard procedure would be the use of the Kalman filter/smoothener [38, 6] which could then be combined with the expectation-maximisation (EM) algorithm [12, 45] to estimate parameters. In this chapter, however, we employ a Bayesian approach sampling directly from the posterior distribution for the rough component, p , without factorising the sampling into forward and backward sweeps.

4.5.1 Path Sampling

The log likelihood functional for the missing data induced by the statistical model (4.2) can be written as follows:

$$\log \mathcal{L}_{IT}(p) = -\frac{1}{2\sigma^2} \sum_{l=0}^N \|\Delta X_l - \Theta A(X_l) \Delta t\|_R^2 + \text{const.} \quad (4.25)$$

We will apply this in the case (4.24) which is a specific instance of (4.2).

One way to sample from this likelihood $\mathcal{L}_{IT}(p)$ for rough paths $\{p_i\}_{i=0}^N$ is via the Langevin equation (see section 6.5.2 in [52]) and, in general, we expect this to be effective in view of the high dimensionality of p . However, when p is Gaussian it is

possible to generate independent samples, and we explain how this may be implemented below.

The Langevin equation is:

$$\frac{dp}{ds} = \nabla_p \log \mathcal{L}_{IT}(p) + \sqrt{2} \frac{dW_s}{ds} \quad (4.26)$$

We explain how the exact sampler (4.29) is derived. The Langevin equation used to sample from the distribution of p (given drift parameters and σ) is:

$$\frac{dp}{ds} = P_{\text{mat}} p + Q(q) + \sqrt{2} \frac{dW}{ds} \quad (4.27)$$

Here, W consists of N independent standard white noise processes and $p = p(s)$ is thought of as a function

$$p : [0, \infty) \rightarrow \mathbb{R}^N,$$

and the form of the derivative $\nabla_p \log \mathcal{L}_{IT}(p)$ employed here will be derived shortly. This equation is continuous in time but discrete in space. Given that the derivative of $\log \mathcal{L}_{IT}$ is linear in the p_i , (4.27) is recognised as an Ornstein-Uhlenbeck process, so that the equilibrium measure is expressible as follows:

$$p \sim \mathcal{N}(-P_{\text{mat}}^{-1}Q(q), -P_{\text{mat}}^{-1}) \quad (4.28)$$

Given a computer-generated pseudo-random i.i.d. sequence of normally distributed random variables, $\{\xi_n\}$, one can generate independent samples with the desired distribution, if the root of the covariance matrix is available, simply by setting:

$$p_n = -P_{\text{mat}}^{-1}Q_{\text{mat}}q + \sqrt{-P_{\text{mat}}^{-1}}\xi_n.$$

As noted above, $-P_{\text{mat}}^{-1}$ is positive definite symmetric. We may thus compute the Cholesky factorisation $U^T U = -P_{\text{mat}}$ and use the following observation which yields

$$\begin{aligned} \mathbb{E} \left(U^{-1} \xi (U^{-1} \xi)^T \right) &= U^{-1} I U^{-T} \\ &= U^{-1} U^{-T} \\ &= -P_{\text{mat}}^{-1} \end{aligned}$$

as desired.

The suggested sampler for p -paths is then

$$p_n = -P_{\text{mat}}^{-1}Q(q) + U^{-1}\xi_n. \quad (4.29)$$

Since a Cholesky factorisation of P_{mat} is an efficient way to compute the mean, the application of U^{-1} is just a backsubstitution using the already computed Cholesky factor.

A cautionary note from Trefethen ([58], p.177) shows that while solving the linear system for P^{-1} is backward stable, the computation of the factor U is not forward-stable, i.e. the errors in U might be large for a generic positive definite matrix. In our case, P is very well-conditioned (Gershgorin yields an upper bound for its condition number with respect to the 2-norm of $\kappa(P) < 3 + \mathcal{O}(\Delta t)$) so that we expect U to be computed accurately. Employing a combination of Theorem 10.5 for stability and Theorem 10.8 for conditioning of the Cholesky factor from [33] this can be substantiated.

Computation of the derivative $\nabla_p \log \mathcal{L}_{IT}$

Now we compute derivatives of the approximate likelihood \mathcal{L}_{IT} needed for a Langevin sampler of the missing path p and for the resulting exact sampler (4.29). We have

$$\begin{aligned} -\sigma^2 \frac{\partial \mathcal{L}}{\partial p_i} &= q_{i+1} \left(\frac{6}{b\Delta t} (\gamma - \Delta t^{-1}) \right) \\ &+ q_i \left(-(1 + \Delta ta) \frac{6}{b\Delta t} (\gamma - \frac{1}{\Delta t}) - \Delta t D (2\Delta t^{-1} - 4\gamma) - 6b^{-1} \Delta t^{-2} \right) \\ &+ q_{i-1} \left((1 + \Delta ta) 6b^{-1} \Delta t^{-2} - 4D \right) \\ &+ p_{i+1} (2\Delta t^{-1} - 4\gamma) \\ &+ p_i (6(\Delta t^{-1} - \gamma) - (2\Delta t^{-1} - 4\gamma)(1 + \Delta t\gamma) + 4\Delta t^{-1}) + p_{i-1} (2\Delta t^{-1} - 4\gamma) \end{aligned}$$

at inner points $0 < i < N$. At the boundary points one gets:

$$\begin{aligned} -\sigma^2 \frac{\partial \mathcal{L}}{\partial p_0} &= q_0 \left(-(1 + \Delta ta) (6b^{-1} \Delta t^{-1} \gamma - 6b^{-1} \Delta t^{-2}) - 2D + 4\gamma D \Delta t \right) \\ &+ q_1 (6b^{-1} \Delta t^{-1} \gamma - 6b^{-1} \Delta t^{-2}) \\ &+ p_0 \left(-\Delta tb (-6b^{-1} \Delta t^{-2} + 6b^{-1} \Delta t^{-1} \gamma) - (1 + \Delta \gamma) (2\Delta t^{-1} - 4\gamma) \right) \\ &+ p_1 (2\Delta t^{-1} - 4\gamma) \end{aligned}$$

And for $i = N$:

$$-\sigma^2 \frac{\partial \mathcal{L}}{\partial p_N} = q_{N-1} ((1 + \Delta t a) 6b^{-1} \Delta t^{-2} - 4D) + q_N (-6b^{-1} \Delta t^{-2}) \\ + p_{N-1} (6\Delta t^{-1} - (1 + \Delta t \gamma) 4\Delta t^{-1}) + p_N (4\Delta t^{-1})$$

These derivatives can be expressed using a tridiagonal, negative definite matrix P_{mat} with highest order stencil $-1 \ -4 \ -1$ acting on the p -vector plus a possibly nonlinear contribution $Q(q)$ acting on the q -vector only. The gradient of \mathcal{L}_{IT} can then be written as claimed:

$$\nabla_p \log \mathcal{L}_{IT}(q, p) = P_{\text{mat}} p + Q(q).$$

4.5.2 Estimating diffusion coefficient and missing path

The approximate likelihood $\mathcal{L}_{IT}(P, Q | \sigma, \Theta)$ can be used to estimate both the missing path p and the diffusion coefficient σ for our Model Problems I, II and III.

In order to estimate σ , the derivative of the log likelihood

$$\log \mathcal{L}_{IT}(\sigma) = \log \mathcal{L}_{IT}(P, Q | \sigma, \Theta) + \log \left(\frac{p_0(\Theta, \sigma)}{\mathcal{L}(P, Q, \Theta)} \right)$$

(where priors $p_0(\Theta, \sigma)$ are assumed to be given) with respect to σ is computed:

$$\frac{\partial}{\partial \sigma} \log \mathcal{L}_{IT} = -\frac{2N}{\sigma} + \frac{1}{\sigma^3} Z + \frac{\partial}{\partial \sigma} \log(p_0(\Theta, \sigma)).$$

Here, we have used the abbreviation

$$Z := \sum_{p=0}^{N-1} \left\| \left(\begin{pmatrix} Q_{p+1} \\ P_{p+1} \end{pmatrix} - \begin{pmatrix} Q_p \\ P_p \end{pmatrix} - \Delta t \begin{pmatrix} P_p \\ -f(Q_n) - \gamma P_p \end{pmatrix} \right) \right\|_R^2.$$

In this case no prior distribution was felt necessary in this example, as when $N \rightarrow \infty$ its importance would diminish rapidly. Thus we set $p_0 \equiv 1$. The resulting maximum likelihood estimator is:

$$\widehat{\sigma^2} = \frac{Z}{2N\Delta t} \quad (4.30)$$

Instead of providing just the maximum of the likelihood it may be more desirable to sample from the distribution of σ given observations p and q and the drift parameters.

As the derivative of the log-likelihood conditional on these observations is available we can write a Langevin type sampler for this distribution in the following form:

$$\begin{aligned} d\sigma &= \frac{\partial \mathcal{L}_{IT}}{\partial \sigma} ds + \sqrt{2}dW \\ &= \left(-\frac{2N}{\sigma} + \frac{1}{\sigma^3}Z \right) ds + \sqrt{2}dW \end{aligned}$$

Empirically, the singularity at $\sigma = 0$ is seen to be more amenable to numerical solution if the transformation $\zeta(\sigma) = \sigma^4$ is used. Using the Itô formula, this yields the Langevin sampler:

$$d\zeta = \left((12 - 8N)\sqrt{\zeta} + 4Z \right) ds + 4\sqrt{2}\zeta^{\frac{3}{4}}dW. \quad (4.31)$$

A simple explicit Euler-Maruyama discretisation in s is used to simulate paths for this SDE.

This Langevin-type sampler (4.31) can then be alternated in a Systematic-Scan Gibbs Sampler (as described on p.130 of [37]) using N_{Gibbs} iterations with the direct sampler for the paths, (4.29). This yields estimates of the missing path and the diffusion coefficient, where the latter is estimated by averaging over the N_{Gibbs} samples of the Gibbs sampler. We illustrate this with an example. For Model Problem I we use the following parameters:

$$\sigma = 1 \quad T \in \{10, 100\} \quad \Delta t \in \{0.1, 0.01\} \quad N_{\text{Gibbs}} = 10$$

The sample paths used for the fitting are generated from exact samples using (4.13) and the resulting plot is given in Figure 4.2 where the first row corresponds to the behaviour when complete observations are available and the second row corresponds to only the smooth component being observed. For Model Problem II we use the following parameters:

$$\begin{aligned} \sigma &= 1 & D &= 4 & \gamma &= 0.5 \\ T &\in \{10, 100\} & \Delta t &\in \{0.02, 0.002\} & N_{\text{Gibbs}} &= 10 \end{aligned}$$

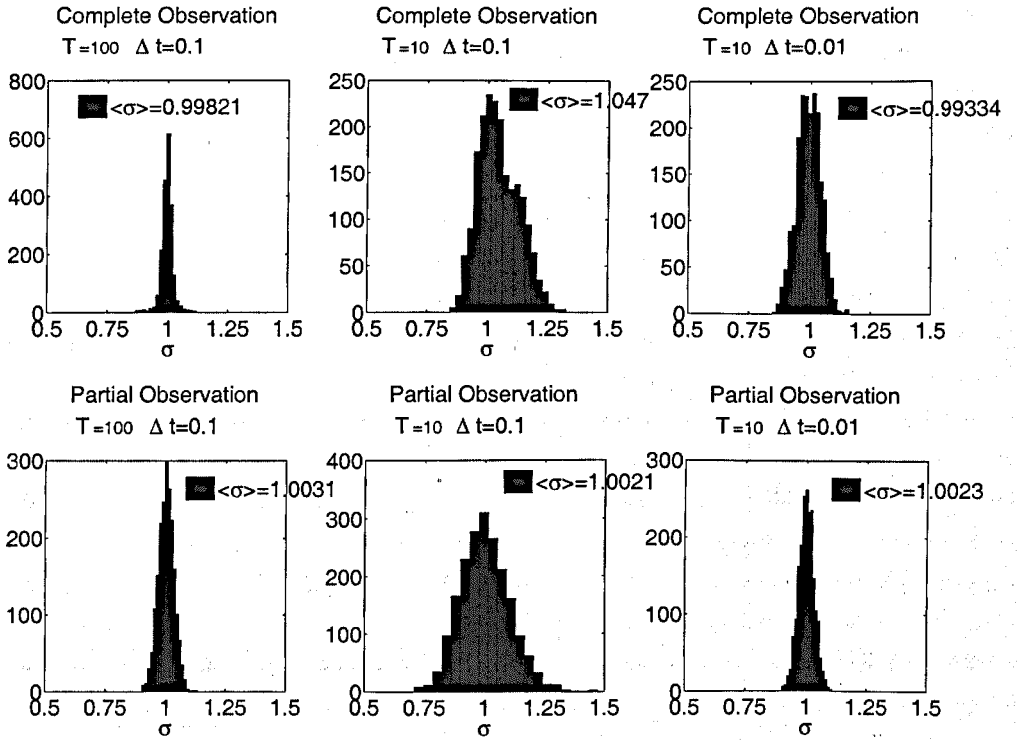


Figure 4.2: Estimates of σ using the \mathcal{L}_{IT} Model for Model Problem I. Top row: fully observed process; bottom row: partially observed process.

The sample paths used for the fitting are generated using a subsampled Euler-Maruyama method with temporal grid $\frac{\Delta t}{k}$ where $k = 30$. This experiment results in the plot given in Figure 4.3.

It appears from these figures that the estimator for this joint problem performs well for Model Problems I and II for Δt sufficiently small and T sufficiently large. A more careful investigation of the convergence properties is postponed to section 6 when drift estimation will be incorporated in the procedure.

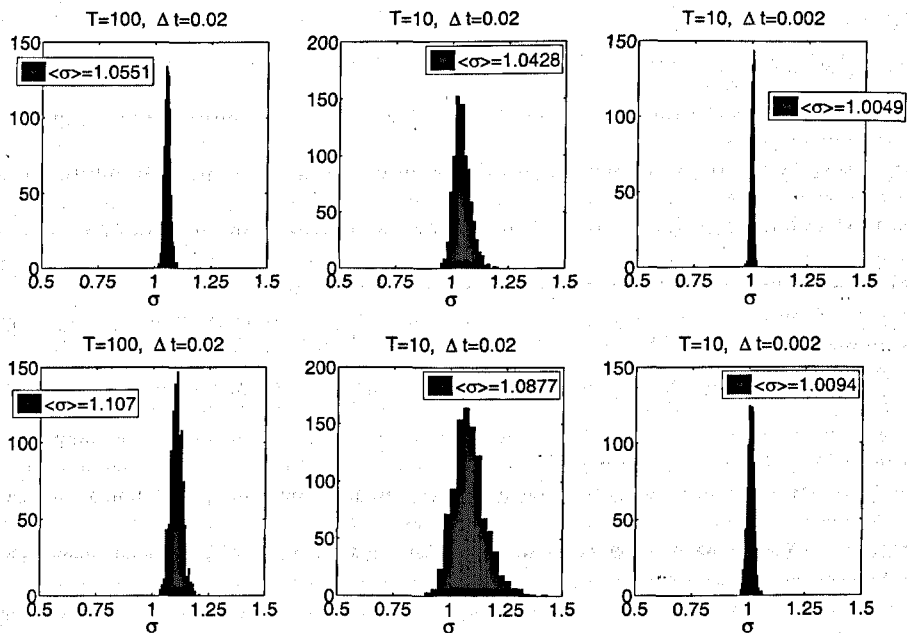


Figure 4.3: Estimates of σ using the \mathcal{L}_{IT} Model for Model Problem II. Top row: fully observed process; bottom row: partially observed process.

4.6 Drift Estimation

4.6.1 Overview

With the approximate likelihoods \mathcal{L}_E and \mathcal{L}_{IT} in place, the question arises which of these should be used to estimate the drift parameters. Using Model Problem II we numerically observe that an \mathcal{L}_E based maximum likelihood estimator performs well. In contrast, ill-conditioning due to hypoellipticity leads to error amplification and affects the performance of the \mathcal{L}_{IT} based estimator. The ill-conditioning is made explicit using asymptotic singularity of the diffusion matrix RR^{-1} .

Alternatively, the estimator (4.9) suggested by Le Breton and Musiela can be used, but this is inappropriate if a harmonic oscillator fit is sought, as it means that all entries of Θ must be estimated and known entries of Θ cannot be fixed a priori. While it is possible to use a cut-back version of this estimator applying it to only those rows of Θ whose entries need to be estimated, it is unclear how to obtain an approximate likelihood corresponding to this estimator that is amenable to Langevin sampling of the drift parameters and – at the same time – avoids the error amplification observed in the \mathcal{L}_{IT} -based case.

Hence, since the \mathcal{L}_E -based estimators also cover Model Problem III, and since they are amenable to Langevin sampling, they are our choice for estimating drift parameters.

4.6.2 Drift parameters from \mathcal{L}_E

In order to simplify analysis, we illustrate the estimator using mainly the Model Problem II, (4.14). Nonetheless, we start from the more general equation (4.11) for which the Euler statistical model is given as follows:

$$\begin{cases} Q_{n+1} = Q_n + \Delta t P_n \\ P_{n+1} = P_n + \Delta t \sum_{i=1}^c D_i f_i(Q_n) - \Delta t \gamma P_n + \sqrt{\Delta t} \sigma \xi_n \end{cases} \quad (4.32)$$

Here, we assume that the force functions $\{f_i\}_{i=1}^c$ are prescaled by parameters $D_i \in \mathbb{R}$. The likelihood functional in this case is given by:

$$\mathcal{L}_E(\gamma, D|Q, P, \sigma) \propto \frac{1}{\sqrt{2\pi\Delta t\sigma^2}^N} \exp\left(-\sum_{n=0}^{N-1} \frac{(\Delta P_n - \Delta t \sum_{i=1}^c D_i f_i(Q_n) + \Delta t \gamma P_n)^2}{2\Delta t\sigma^2}\right) \quad (4.33)$$

Differentiating this likelihood with respect to the parameters $\{D_i\}_{i=1}^c$ and γ and equating to zero yields a linear system of equations which we denote by

$$M_E \begin{bmatrix} D_1 \\ \vdots \\ D_c \\ \gamma \end{bmatrix} = b_E. \quad (4.34)$$

In the harmonic oscillator case of Model Problem II, where $c = 1$ and $f_1(q) = -Dq$ we obtain the following linear system:

$$\begin{bmatrix} \sum_{n=0}^{N-1} \Delta t Q_n^2 & \sum_{n=0}^{N-1} \Delta t Q_n P_n \\ \sum_{n=0}^{N-1} \Delta t Q_n P_n & \sum_{n=0}^{N-1} \Delta t P_n^2 \end{bmatrix} \begin{bmatrix} \hat{D} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} -\sum_{n=0}^{N-1} Q_n \Delta P_n \\ -\sum_{n=0}^{N-1} P_n \Delta P_n \end{bmatrix} \quad (4.35)$$

The continuum limit for $\Delta t \rightarrow 0$ with $N\Delta t = T$ of this system is simply:

$$\begin{bmatrix} \int_0^T q(t)^2 dt & \int_0^T p(t)q(t) dt \\ \int_0^T p(t)q(t) dt & \int_0^T p(t)^2 dt \end{bmatrix} \begin{bmatrix} \hat{D} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} -\int_0^T q(t) dp_t \\ -\int_0^T p(t) dp_t \end{bmatrix}$$

This corresponds to the estimator of D and γ alone given by (4.9). Casting aside issues about the discretisation error (finite Δt), the proof of asymptotic consistency given in [4] still applies to this estimator in the linear case.

Using the same likelihood, \mathcal{L}_E , a Langevin sampler can also be used for the drift parameters. Since the resulting distribution for Θ is Gaussian, direct sampling can be used in the spirit of subsection 4.5.1:

$$\hat{\Theta} \sim \mathcal{N}(M_E^{-1}b_E, M_E^{-1}) \quad (4.36)$$

4.6.3 Drift parameters from \mathcal{L}_{IT}

As the approximate model based on \mathcal{L}_{IT} is observed to resolve the difficulty with estimating σ for hidden p -paths, it is interesting to see whether it can also be used to estimate the drift parameters.

The log-likelihood function is given by (4.25). To illustrate the problems arising from the use of \mathcal{L}_{IT} we use Model Problem II, so that (4.25) becomes

$$\log \mathcal{L}_{IT}(\Theta) = \frac{1}{2\sigma^2\Delta t} \sum_{n=0}^{N-1} \|(X_{n+1} - X_n - \Delta t\Theta A(X_n))\|_R^2 + \text{const} \quad (4.37)$$

where $R = \begin{bmatrix} \frac{\Delta t}{\sqrt{12}} & \frac{\Delta t}{2} \\ 0 & 1 \end{bmatrix}$, irrelevant constants have been omitted and we have

$$A \left(\begin{bmatrix} Q_n \\ P_n \end{bmatrix} \right) = \begin{bmatrix} Q_n \\ P_n \end{bmatrix}, \quad \theta = \begin{bmatrix} 0 & 1 \\ -D & -\gamma \end{bmatrix}.$$

In order to obtain a maximum likelihood estimator from this, we take the derivative with respect to the parameters D and γ and equate to zero. This yields the following linear system:

$$\begin{bmatrix} \sum_n Q_n^2 \Delta t & \sum_n P_n Q_n \Delta t \\ \sum_n P_n Q_n \Delta t & \sum_n P_n^2 \Delta t \end{bmatrix} \begin{bmatrix} \hat{D} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} -\sum_n Q_n \Delta P_n \\ -\sum_n P_n \Delta P_n \end{bmatrix} + \begin{bmatrix} \sum_n \frac{3}{2} Q_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \\ \sum_n \frac{3}{2} P_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \end{bmatrix} \quad (4.38)$$

Comparing this linear system to the successful estimator (4.34) we note the presence of an additional term on the right hand side. This term leads to the failure of the above estimator.

4.6.4 Numerical Check: Drift

There are two factors influencing convergence: T and Δt . To illustrate their influence, consider the following series of numerical tests. All of the tests share these parameters:

$$D = 4 \quad \gamma = 0.5 \quad \sigma = 0.5 \quad k = 30$$

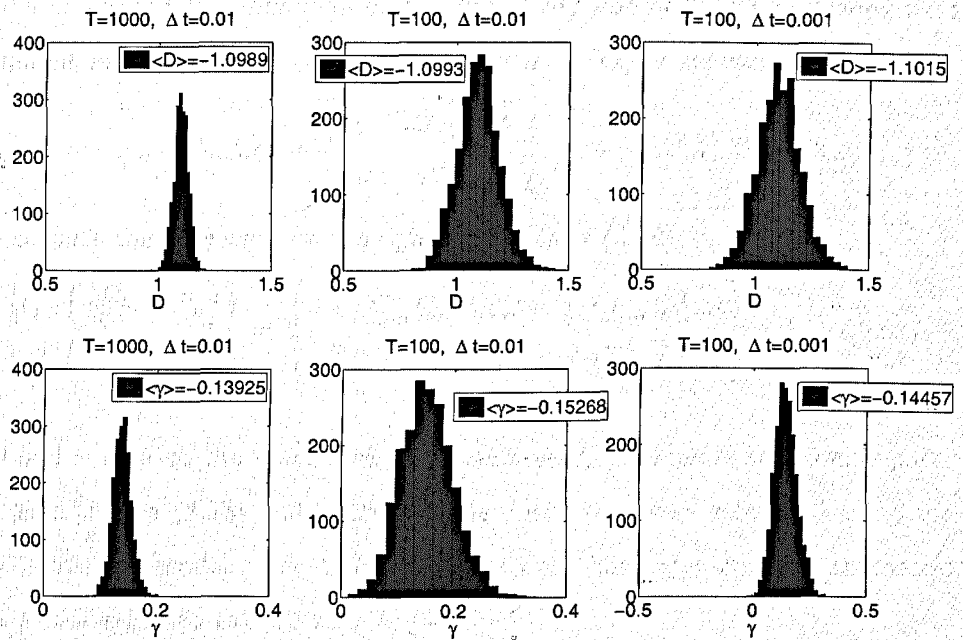


Figure 4.4: Drift estimation for Model Problem II, using \mathcal{L}_{IT}

Data for the tests are again generated using an Euler-Maruyama method on a finer temporal grid with resolution $\Delta t/k$. In the plot given in Figure 4.4 the top row contains histograms for the drift parameter D whereas the second row contains histograms for the drift parameter γ in any case using the full sample path for inference. It is clear from these experiments summarised in Figure 4.4 that both D and γ are grossly underestimated.

4.6.5 Why the Model fails for the drift parameters

The key is to analyse the error term on the right hand side of (4.38) comparing it to the consistent estimator (4.34). Using the 2nd order Itô-Taylor approximation

$$X_{n+1} = X_n + \Delta t A X_n + \begin{bmatrix} 1 & 0 \\ -\gamma & 1 \end{bmatrix} R \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \frac{1}{2} \Delta t^2 A^2 X_n + \mathcal{O}(\Delta t^{\frac{5}{2}})$$

we can compute the error term on the right hand side of (4.38):

$$\begin{bmatrix} \sum_n \frac{3}{2} Q_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \\ \sum_n \frac{3}{2} P_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \end{bmatrix} = \begin{bmatrix} -\frac{3}{4} \gamma \sum_n Q_n P_n \Delta t - \frac{3}{4} D \sum_n Q_n^2 \Delta t \\ -\frac{3}{4} D \sum_n Q_n P_n \Delta t - \frac{3}{4} \gamma \sum_n P_n^2 \Delta t \end{bmatrix} + I_s + \mathcal{O}(\Delta t). \quad (4.39)$$

Here, D and γ refer to the exact drift parameters used to *generate* the sample path, whereas \hat{D} and $\hat{\gamma}$ in (4.38) and (4.39) are the drift parameters estimated using the improved statistical model. The term I_s on the right hand side contains stochastic integrals whose expected value is zero.

As the mean error terms can be written in terms of the matrix elements themselves, (4.39) can be substituted in (4.38) to obtain:

$$\mathbb{E} \hat{D} = \frac{1}{4} D + \mathcal{O}(\Delta t) \quad (4.40)$$

$$\mathbb{E} \hat{\gamma} = \frac{1}{4} \gamma + \mathcal{O}(\Delta t). \quad (4.41)$$

This seems to be corroborated by the numerical tests.

4.6.6 Analysis of Drift Estimation Failure

In order to make the ill-conditioning whose effects were exhibited in subsection 4.6.3 more explicit, a more general analysis is attempted.

It was seen above that the Euler statistical model does not allow for successful estimation of σ in the case of hidden p -path whereas the improved model delivers correct estimates. On the other hand, the latter model delivers incorrect estimates of the drift parameters even if the complete path is known.

Essentially, this stems from the fact that the correlation matrix RR^{-1} is not factored out of the drift parameter estimation as it should be. This leads to $\mathcal{O}(\Delta t)$ errors of Θ in the first (q -) row of the equations (where you implicitly assume exact knowledge of the upper row of the drift matrix rather than allowing $\mathcal{O}(\Delta t)$ errors) to be amplified by the Δt -dependent coefficients of RR^{-1} to $\mathcal{O}(1)$ errors of Θ in the second row of equations.

To analyse this failure more carefully, we derive an estimator which estimates *all* entries of Θ and in the process this interaction of ill-conditioning as $\Delta t \rightarrow 0$ and the suppression of small errors in some matrix entries is demonstrable.

Statistical model

Assume, like above, that some higher order estimate of the correlation matrix $R(\Delta t; \bar{\Theta})$ is given, where $\bar{\Theta}$ is written to indicate that it is supposed to depend only on those entries of Θ that are known and not to be estimated. (The matrix used in the construction of \mathcal{L}_{IT} is such a case.) The point of the following calculation, however, is that this matrix will drop out if all of Θ is to be estimated and so the end result of this calculation implies that the dependence of R on Θ is irrelevant if all of Θ is estimated. The only requirement on $R(\Delta t; \Theta)$ is that it be invertible.

Using the statistical model (4.2) we have again:

$$X_{n+1} = X_n + \Delta t \Theta A(X_n) + R(\Delta t; \bar{\Theta}) \xi \quad (4.42)$$

where ξ is now an \mathbb{R}^k -valued normally distributed random variable providing d independent samples from $\mathcal{N}(0, 1)$.

The log likelihood functional can now be written as follows:

$$\mathcal{L} = \frac{1}{2} \sum_{n=0}^N \|R^{-1}(\Delta t; \bar{\Theta}) (X_{n+1} - X_n - \Delta t \Theta A(X_n))\|^2 \quad (4.43)$$

The derivative of \mathcal{L} is a linear function of Θ which satisfies

$$\mathcal{L}(\Theta + \Delta\Theta) = \mathcal{L}(\Theta) + B : \Delta\Theta + \mathcal{O}(\Delta\Theta^2)$$

for the matrix B which is the Riesz-representative of the linear functional

$$\frac{\partial \mathcal{L}}{\partial \theta} : \mathbb{R}^{k \times N} \longrightarrow \mathbb{R}.$$

Expanding (4.43) about Θ yields the following expression for the action of the derivative of the log-likelihood function:

$$\frac{\partial \mathcal{L}}{\partial \Theta}(\Delta\Theta) = -\Delta t \sum_n ((RR^T)^{-1}(X_{n+1} - X_n - \Delta t \Theta A(X_n))) \otimes A(X_n) : \Delta\Theta \quad (4.44)$$

Crucially, if all variations $\Delta\Theta$ are permissible, a necessary condition for the Θ maximising the likelihood is for the matrix on the left hand side of the matrix inner product to be identically zero (this corresponds to R being independent of Θ). In this case, the necessary condition can be written as

$$0 = -\Delta t \sum_p ((RR^T)^{-1}(\Delta X_n - \Delta t \Theta A(X_n))) \otimes A(X_n)$$

where $\Delta X_n = X_{n+1} - X_n$ has been used to simplify notation. Crucially, this can be premultiplied by RR^T – this is the factoring out of the noise model mentioned earlier –

$$\Delta t \sum_p (\Theta A(x_p)) \otimes A(x_p) = -\sum_p \Delta x_p \otimes A(x_p)$$

so that the estimate $\hat{\Theta}$ is given by

$$\hat{\Theta} = \left(\sum_p \frac{-\Delta x_p}{\Delta t} \otimes A(x_p) \right) \left(\sum_p A(x_p) \otimes A(x_p) \right)^{-1}. \quad (4.45)$$

This is but the estimator suggested by Le Breton and Musiela in [4]. They observe it to be asymptotically consistent in the case of linear A and it can be shown to have a bias of order $\mathcal{O}(\Delta t)$ using an Itô-Taylor expansion.

The crucial step in going from (4.44) to (4.45) is that *all* variations $\Delta\Theta$ in (4.44) are permissible so that the matrix to the left of that matrix inner product is identically

zero. If only some entries of Θ are to be estimated (as in the harmonic oscillator example above), only those $\Delta\Theta$ corresponding to variations in the components to be estimated are permissible. To elucidate this issue further, let us rewrite (4.44) using the symmetry of RR^T as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta}(\Delta\Theta) = -\Delta t \sum_n (X_{n+1} - X_n - \Delta t \Theta A(X_n)) \otimes A(X_n) : (RR^T)^{-1} \Delta\Theta \quad (4.46)$$

It is clear from this expression that the direction in which the matrix expression on the left of the inner product must be zero changes as a function of Δt . Whether this change leads to amplification of errors or merely to lower order perturbations depends on the exact entries to be estimated. It is the interaction of the scaling in the matrix R and the choice of parameters to be estimated that causes the observed $\mathcal{O}(1)$ errors.

4.6.7 Conclusion for Drift Estimation

It has been observed numerically that the likelihood \mathcal{L}_E associated with an Euler model for the SDE (4.1) yields asymptotically consistent Langevin and maximum likelihood estimators for Model Problem II. For the case of continuous time the proof of asymptotic consistency in the limit $T \rightarrow \infty$ given in [4] can be adapted in the linear case (i.e. $A = id$) and it would be expected to carry over to the discretised problem in the limit $\Delta t \rightarrow 0$ and $N\Delta t \rightarrow \infty$.

While it is aesthetically desirable to base the estimation of all parameters as well as the missing data on the same approximation \mathcal{L}_{IT} of the true likelihood \mathcal{L} , and although this approximation was found to work well for the estimation of missing data and the diffusion coefficient, it does not work for the drift parameters.

It is possible to trace this failure to the fact that only the second row of Θ is estimated where $\mathcal{O}(\Delta t)$ errors in the first row get amplified to $\mathcal{O}(1)$ errors in the second row. Estimating all entries of Θ , while being outside the specification of the problem under consideration, also yields $\mathcal{O}(1)$ errors if \mathcal{L}_{IT} is used and so does not remedy the problem. This problem is not shared by the discretised version of the diffusion independent estimator (4.9), but this is not a maximum likelihood estimator for \mathcal{L}_{IT} .

In summary, for the purposes of fitting our model problems to observed data we

employ the Euler statistical model (4.33) for the drift parameters.

4.7 The Gibbs Loop

In this section, we combine the insights obtained in previous sections to formulate an effective algorithm to fit hypoelliptic diffusions to partial observations of data at discrete times. We apply a deterministic scan Gibbs sampler alternating between missing data, drift parameters and diffusion parameters. Subsection 4.7.1 describes the approach in the general case, when applied to (4.1), whereas subsection 4.7.2 describes the application to Model Problem III.

4.7.1 Overview

In this section, the estimators for the hidden rough path V , the covariance $\Gamma\Gamma^T$ and the those rows of the drift parameters Θ which are to be estimated are combined in a Gibbs sampler. Given a likelihood $\mathcal{L}(U, V|\Theta, \Gamma\Gamma^T)$, a prior $p_0(\Theta, \Gamma\Gamma^T)$ and observation U , a Systematic Scan Gibbs Sampler would normally work as follows:

1. Sample V from $\mathcal{L}(V|U, \Theta, \Gamma\Gamma^T)$.
2. Sample Θ from $\mathcal{L}(\Theta|U, V, \Gamma\Gamma^T)$.
3. Sample $\Gamma\Gamma^T$ from $\mathcal{L}(\Gamma\Gamma^T|U, V, \Theta)$.
4. Restart from step 1 unless sufficiently equilibrated.

Of course, the exact likelihood for the problem at hand is unavailable and thus approximate likelihoods are chosen. Exactly which approximations are chosen depends on the problem at hand. We have outlined how to construct \mathcal{L}_{IT} approximations to estimate V and $\Gamma\Gamma^T$ by propagating the highest order noise to every row and \mathcal{L}_E approximations for the drift parameter estimation. Numerical and analytical evidence indicates that these approximations work well.

The algorithm to be put in practice thus reads:

1. Sample V from $\mathcal{L}_{IT}(V|U, \Theta, \sigma)$.
2. Sample Θ from $\mathcal{L}_E(\Theta|U, V, \sigma)$.
3. Sample σ from $\mathcal{L}_{IT}(\sigma|U, V, \Theta)$.
4. Restart from step 1 unless sufficiently equilibrated.

In practice, we find that for Model Problem II and III, equilibration is fast. Furthermore, convergence of the estimates to the true parameter values is observed numerically for Model Problems II and III with $\mathcal{O}(\Delta t)$ discretisation errors and $\mathcal{O}(\frac{1}{T})$ truncation errors if the sample paths do not start in the equilibrium measure. The overall bias is therefore of order $\mathcal{O}(\Delta t + \frac{1}{T})$ and the observed variance is of order $\mathcal{O}(\frac{1}{T})$. We now show this in detail.

4.7.2 The Algorithm

The proposed algorithm will be illustrated using Model Problem III.

Algorithm 2. Given observations $q_i, i = 1, \dots, N$, the initial p -path is obtained using numerical differentiation:

$$p_i^{(0)} = \frac{\Delta q_i}{\Delta t}. \quad (4.47)$$

The initial drift parameter estimate is just set to zero: $\{D_j^{(0)}\}_{j=1}^c = 0, \gamma^{(0)} = 0$. Then start the Gibbs loop:

For $k = 1, \dots, N_{\text{Gibbs}}$:

1. Estimate the drift parameters $\gamma^{(k)}$ and $\{D_j^{(k)}\}_{j=1}^c$ using sampling from \mathcal{L}_E given $\{p_i^{(k-1)}\}_{i=0}^N$ via (4.36).
2. Estimate the diffusivity $\sigma^{(k)}$ using the Langevin sampler (4.31) based on \mathcal{L}_{IT} given $\{p_i^{(k-1)}\}_{i=0}^N$ and $\gamma^{(k)}, \{D_j^{(k)}\}_{j=1}^c$.
3. Get an independent sample of the p -path, $\{p_i^{(k)}\}_{i=0}^N$ using (4.29) derived from \mathcal{L}_{IT} given parameters $\gamma^{(k)}, \{D_j^{(k)}\}_{j=1}^c$ and $\sigma^{(k)}$.

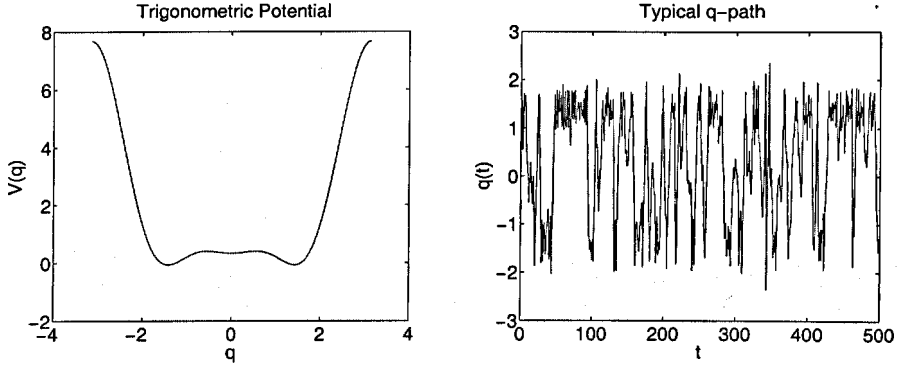


Figure 4.5: Typical sample path for Model Problem III, $T = 500$

This algorithm is tested numerically where sample paths of (4.15) are generated using a sub-sampled Euler-Murayama approximation of the SDE. The data is generated using a timestep that is smaller than the observation time step by a factor of either $k = 30$ or $k = 60$. Comparing the results for these two and other non-reported cases, they are found not to depend on the rate of subsampling, k , if this is chosen large enough. The parameters used for these simulations are as follows:

$$\begin{aligned}
 D_0 = 1 \quad D_1 = -8 \quad D_2 = 8 \quad \gamma = 0.5 \quad \sigma = 0.7 \\
 T = 500 \quad \Delta t \in \left\{ \frac{1}{2}, \dots, \frac{1}{128} \right\} \quad N_{\text{Gibbs}} = 10
 \end{aligned}$$

The trigonometric potential resulting from this choice of drift parameters is depicted on the left of Figure 4.5 and a typical sample path is given on the right side of Figure 4.5. It should be noted that all sample paths are started at $(q, p) = (1, 1)$. As the potential is inspired by dihedral angle potentials used in molecular dynamics it seems appropriate that σ is chosen such that metastability occurs. This can be observed in the typical q -path given in Figure 4.5.

Using up to 64000 sample paths we obtain estimates of the drift parameters by averaging over the latter half of $N_{\text{Gibbs}} = 50$ Gibbs iterations. We label these as $\langle \widehat{D}_i \rangle$ and $\langle \widehat{\gamma} \rangle$. We then compute their deviation from the true values, $\Delta D_i = \langle \widehat{D}_i \rangle - D_i$ and plot ΔD_i and $\Delta \gamma$ versus Δt in a doubly logarithmic plot given in Figure 4.6.

A similar plot which is given in Figure 4.7 is obtained for the shorter final time

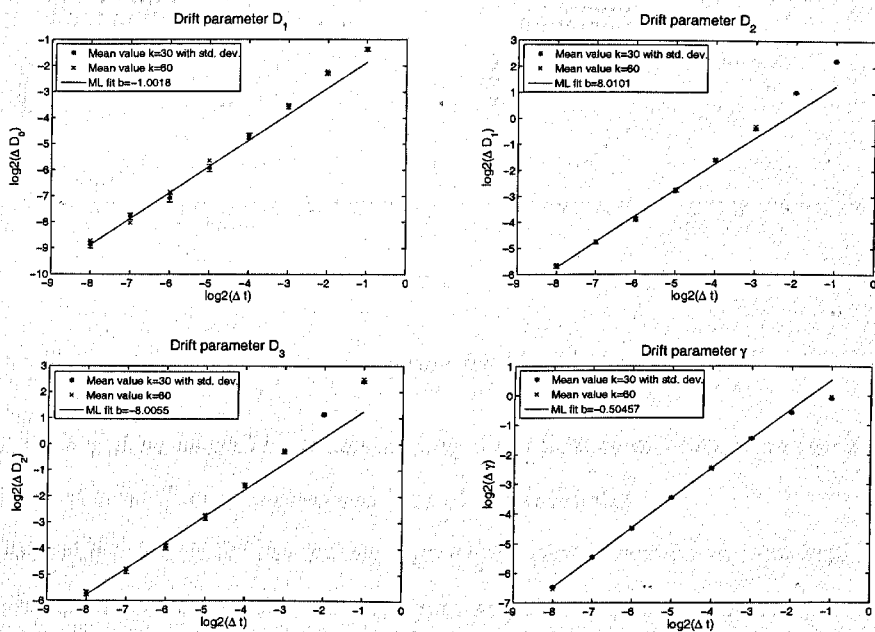


Figure 4.6: Whole loop estimation for Model Problem III: $T = 500$

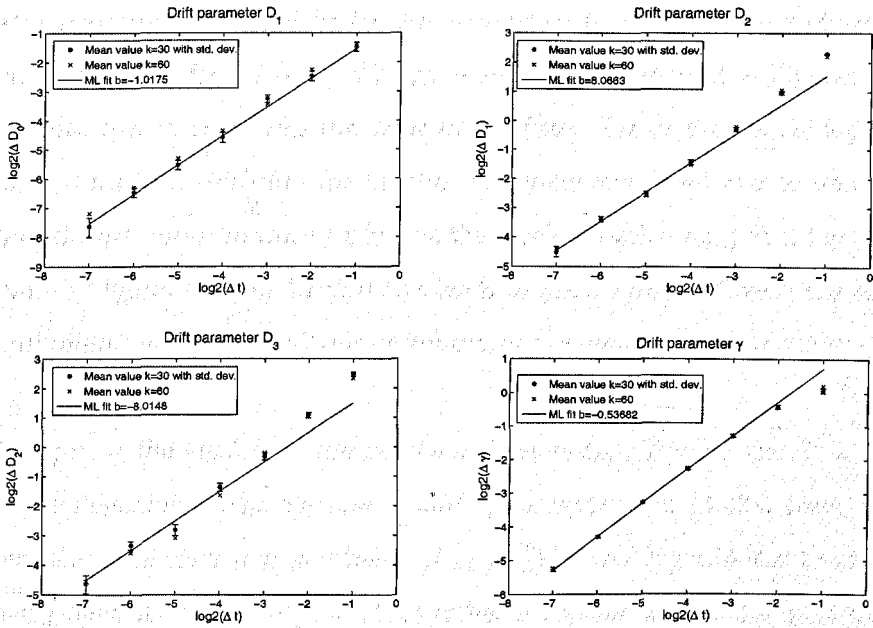


Figure 4.7: Whole loop estimation for Model Problem III: $T = 50$

$T = 50$ which will be helpful in understanding the influence of finite time resolution Δt and finite final time T on the observed bias of the estimators.

A straight line fit for the doubly logarithmic plot is desired to numerically ascertain the order of convergence. First attempts at obtaining such a fit using a standard least squares procedure yield a slope close to 1 indicating $\mathcal{O}(\Delta t)$ errors in the fitted parameters. However, since the Monte Carlo standard deviations around each datapoint get magnified due to the logarithmic transformation, the fact that the apparent variance increases as Δt is decreased has to be taken into account. As the observed transformed standard deviations cannot be assumed to be small in comparison to the observed mean error, a more sophisticated method than the standard least squares fit is suggested.

Given averaged numerically observed parameter estimates y_i and their numerically observed Monte Carlo standard deviations α_i obtained at timesteps Δt_i we fit b and c in the following model:

$$\alpha_i \xi_i = y_i - b - c \Delta t_i. \quad (4.48)$$

Assuming that the errors ξ_i are normally distributed (which is empirically found to be the case) a maximum likelihood fit for the parameters b and c can be performed and yields the asymptotic (for $\Delta t \rightarrow 0$) drift parameter values reported in Figures 4.6 and 4.7. Note that this fit constrains the slope of the fitted line in the doubly logarithmic plot to one. This is to minimise the number of parameters fitted and to improve the accuracy of the extrapolated value b which is the predicted value for y at $\Delta t = 0$. It can be observed in Figures 4.6 and 4.7 that this leads to good agreement with the observed average parameter values y_i , and this corroborates the estimator's bias being of order $\mathcal{O}(\Delta t)$.

Comparing the results for the two final times tested, $T = 50$ and $T = 500$, we find that the deviation of the asymptotic drift parameter (b in (4.48)) from the true parameter value is consistent with it being $\mathcal{O}(\frac{1}{T})$. This error is attributed to all sample paths having been started at $(q, p) = (1, 1)$ rather than from a point sampled from the equilibrium measure.

For the diffusion parameter σ , results analogous to those in Figure 4.6, using the same parameter values, are shown in Figure 4.8 (although that figure displays results for $k = 30$ only). Asymptotic consistency can be observed from this figure with a naive least squares fit yielding a slope of $\mathcal{O}(\Delta t^{0.93})$. This is consistent with an $\mathcal{O}(\Delta t)$ error in the estimated diffusion parameter.

From these considerations it is apparent that the numerical experiments' outcome is consistent with an $\mathcal{O}(\Delta t) + \mathcal{O}(\frac{1}{T})$ bias, making the Algorithm 2 an asymptotically consistent estimator of the drift and diffusion parameters.

4.7.3 Combining MLE and Langevin estimators in a Gibbs Sampler

The Gibbs algorithm alternating between different Langevin samplers as described in section 4.7.1 is suitable whenever it is possible to sample from those likelihoods. It was noted in 4.6.1 that a cut-back version of the estimator described in [4] could be used to estimate parameters for a harmonic oscillator fit but that it was unclear how to convert this maximum likelihood estimator into a sampler for a suitable posterior distribution. There may also be other cases where practical computational considera-

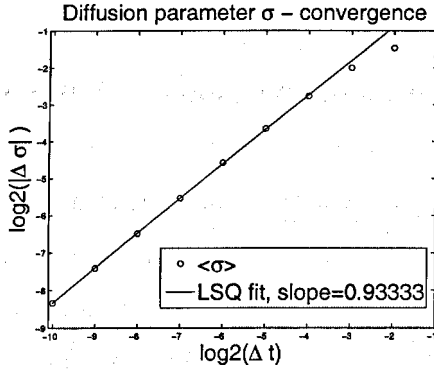


Figure 4.8: Whole loop estimation for Model Problem III: $T = 500$

tions enable maximum likelihood estimates (or good approximations thereof) but *not* the corresponding posterior sampler. In these cases, it is interesting to ask whether combining maximum likelihood estimators and samplers for posterior distribution in a common deterministic scan Gibbs sampler is statistically viable. In this subsection, this question will be answered analytically in the case of a 2d Gaussian example.

The standard deterministic scan Gibbs sampler as described in [37], p. 130, assumes the following setup: For a probability distribution

$$p : \mathbb{R}^n \longrightarrow \mathbb{R}^+$$

where it is assumed that sampling the conditional distributions $p(x_1|x_2, \dots, x_n)$ is possible at low computational cost. The suggested algorithm then is to deterministically cycle through the x_i updating one at a time:

- Provide starting guess for x_1^1, \dots, x_n^1 .
- for $k = 1, 2, \dots$
 1. for $i = 1, 2, \dots, n$

(a) Sample x_i^{k+1} from $p(x_i|x_1^k, \dots, x_{i-1}^k, x_{i+1}^k, \dots, x_n^k)$

It is easy to see that this algorithm leaves the true distribution p invariant and according to [37], geometric convergence can be shown under weak hypotheses.

One variant of the algorithm presented above to sample the missing path, the drift parameters and the diffusion coefficient employs a combination of maximum likelihood estimators for the drift parameters (and possibly the diffusion coefficient) and either a Langevin or a direct sampler for the missing path. In addition, it also uses different approximations to the true probability distribution p for the different estimators, but even if no such approximations were necessary, it still would not be obvious whether such a combination should be expected to yield correct statistics. Such hybrid methods seem not to be routinely used in statistics.

The fact that convergence is observed numerically for such hybrid methods fosters a belief in their usability, so in order to gain some analytical understanding, the 2D Gaussian case is analysed here. Of course, the joint probability distribution to be sampled from in practice will not be Gaussian. However, it is expected to be unimodal (otherwise it is very simple to construct counterexamples) and approximately Gaussian.

Given a probability distribution $p(x, y)$ on \mathbb{R}^2 , the algorithm to be analysed can be written as follows:

- Provide starting guess for x_1, y_1 .
- for $k = 1, 2, \dots$
 1. Sample $y_{k+1} \sim p(y, x_k)$ using the correct marginal distribution.
 2. Sample x_{k+1} using an MLE: $x_{k+1} = \operatorname{argmax}_x p(x|y_{k+1})$

Since the algorithm is translation invariant, it suffices to treat a 2D Gaussian distribution centered at 0:

$$p(x, y) = \frac{1}{2\pi} \sqrt{\det C} \exp \left(-\frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^T C \begin{bmatrix} x \\ y \end{bmatrix} \right) \quad (4.49)$$

where C is a positive definite symmetric matrix with entries c_{ij} which is the inverse of the variance of p . The questions to be answered are:

- Does the algorithm yield the correct expected value? Does this depend on the starting values x_1, y_1 ?

- Does the algorithm produce data with the correct (marginal) variances of x and y ?

The answer to the first question is affirmative as one would expect whereas the second answer is negative and the error is quantified.

Starting with a given intermediary value x_k , y_{k+1} will be distributed according to $p(y|x_k)$, i.e. $y_{k+1} \sim \mathcal{N}\left(-x_k \frac{c_{12}}{c_{22}}\right)$. Moving on, x_{k+1} will then be given as follows:

$$\begin{aligned} x_{k+1} &= \operatorname{argmax}_x \exp\left(-\frac{1}{2}x^2 c_{11} - xy_{k+1}c_{12} - \frac{1}{2}y_{k+1}^2 c_{22}\right) \\ &= \operatorname{argmax}_x \exp\left(-\frac{c_{11}}{2}\left(x + y_{k+1} \frac{c_{22}}{c_{11}}\right)^2 - \frac{1}{2}y_{k+1}^2 c_{22} + \frac{1}{2}y_{k+1}^2 \frac{c_{12}^2}{c_{11}}\right) \\ &= -\frac{c_{12}}{c_{11}}y_{k+1} \end{aligned}$$

So, overall the iteration relation is:

$$x_{k+1} \sim \mathcal{N}\left(x_k \frac{c_{12}^2}{c_{11}c_{22}}, \frac{1}{c_{22}}\right) \quad (4.50)$$

Since C is positive definite symmetric we have that $\det C > 0$ and thus $c_{11}c_{22} > c_{12}^2$. The iteration relation for the expected values is simply $\mathbb{E}x_{k+1} = \frac{c_{12}^2}{c_{11}c_{22}}$, and this is now seen to have an attractive fixed point at $x = 0$ which is the true value. The expected value for y_k behaves accordingly.

In order to answer the second question, the variance of x_{k+1} has to be computed:

$$\begin{aligned} \operatorname{Var}(x_{k+1}) &= \frac{c_{12}^2}{c_{11}^2} \operatorname{Var}(y_{k+1}) \\ &= \frac{c_{12}^2}{c_{11}^2} \left(\frac{c_{12}^2}{c_{22}} \operatorname{Var}(x_k) + \frac{1}{c_{22}} \right) \end{aligned}$$

The iteration equation for the marginal variance for x under this algorithm is thus given as:

$$v_{k+1} = \frac{c_{12}^4}{c_{11}^2 c_{22}^2} v_k + \frac{c_{12}^2}{c_{11}^2 c_{22}}$$

where $v_k = \operatorname{Var}x_k$ was used to simplify notation. Exploiting positive definiteness of C one again finds that this equation has an attractive fixed point at the value:

$$v^* = \frac{c_{22}c_{12}^2}{c_{11}^2 c_{22}^2 - c_{12}^4} \quad (4.51)$$

This should be compared to the true marginal variance of x which can be evaluated from the marginal distribution

$$x \sim \int_{\mathbb{R}} \frac{1}{2\pi} \sqrt{\det C} \exp \left(-\frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^T C \begin{bmatrix} x \\ y \end{bmatrix} \right) dy$$

and is found to be

$$\text{Var}_{\text{true}} = \frac{c_{22}}{c_{11}c_{22} - c_{12}^2}. \quad (4.52)$$

So comparing (4.51) and (4.52) the variance is incorrect by a factor

$$\begin{aligned} \frac{v^*}{\text{Var}_{\text{true}}} &= \frac{c_{12}^2}{c_{11}c_{22} + c_{12}^2} \\ &\leq 1 \end{aligned}$$

It is unsurprising that the variance is underestimated by this semi-deterministic algorithm. This result cautions that Monte-Carlo variances along Markov Chain Monte Carlo simulations should not be used to estimate the posterior variance of the estimated parameters if a hybrid algorithm is used.

Just how much the variance may be underestimated in practice can be observed from the following example. We consider the trigonometric oscillator with linear damping as given in equation (4.15). The following parameters were used:

$$T_f = 500 \quad D_0 = 1 \quad D_1 = -8 \quad D_2 = 8 \quad \gamma = 0.5 \quad \sigma = 0.7$$

In the plot given as figure 4.9 we compare histograms for a certain drift parameter, Θ_2 . Firstly, the distribution of Θ_2 given smooth sample paths q , i.e. $p(\Theta_2|q)$ is shown obtained from repeated experiments using the all-sampling algorithm given in section 4.7.2. This is contrasted with the posterior distribution for one particular, fixed, realisation of q which is obtained using the same all-sampling algorithm. It can be seen that the observed Monte Carlo variance provides a good estimate of the true variance of the drift parameter even though the expected value is subject to a fairly large deviation. Furthermore, this is contrasted with the approximation to the posterior distribution obtained using the same fixed realisation q but this time sampled using a hybrid method

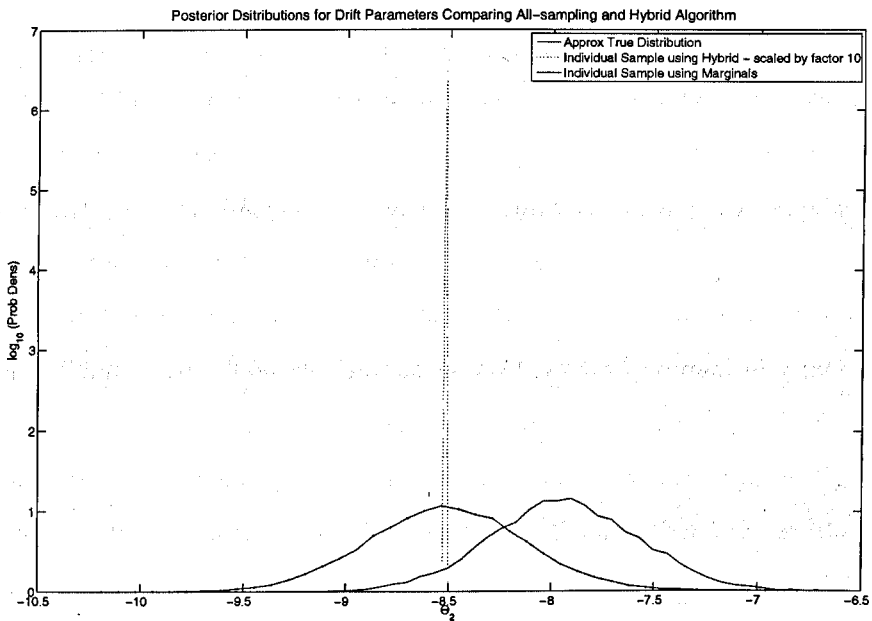


Figure 4.9: Comparing Hybrid and all-sampling Algorithms

with an MLE for the drift parameters Θ and a direct sampler for the missing path p . It is observed that while the expected value is the same as with the all-sampling method, the variance is grossly underestimated. It should be stressed that this example uses *different approximations* to the true likelihood in different steps of the algorithm throughout, furthermore the distributions are not Gaussian so that – strictly speaking – the above analysis does not apply.

This hybrid algorithm takes a middle ground between the EM (Expectation-Maximisation) algorithm where the sampling from the marginal distribution would be replaced by computing the expected value and the Deterministic Scan Gibbs Sampler. A comparison in terms of convergence rates of those two has been carried out by Roberts and Sahu, [56].

4.8 Application to Molecular Conformational Dynamics

As a simple application of fitting hypoelliptic diffusions using partial observations we consider data arising from molecular dynamics simulations of a butane molecule using a simple heat bath approximation. After describing the origin of the data to be fitted, we observe that for small Δt , fitting an elliptic diffusion process is inappropriate as the fitted diffusion coefficient $\hat{\sigma}$ tends to zero as $\Delta t \rightarrow 0$.

By considering the origin of the data we demonstrate that it is natural to fit a hypoelliptic diffusion process which yields convergent results for diminishing inter-sample intervals Δt . Also, stabilisation of the fitted force function $f(q) = \sum_{j=1}^c D_j f_j(q)$ as the number of terms to be included, c , increases, is observed. Thus the hybrid Algorithm 2 is shown to be effective on real data. It is also clear, though, that the resulting fit has only limited predictive capabilities as it fails to fit the invariant measure of the data at all well. However, this is a *modeling* issue which is not central to this chapter.

4.8.1 Molecular Dynamics

The data used for this fitting example are generated using a molecular dynamics (MD) simulation for a single molecule of butane. In order to avoid explicit computations for solvent molecules, several *ad hoc* approximate algorithms have been developed in molecular dynamics. One of the more sweeping approximation that is nonetheless fairly popular, at least as long as electrostatic effects of the solvent can be neglected or treated otherwise, is Langevin dynamics. The butane molecule is modelled as a damped-driven Hamiltonian system of the form

$$\ddot{x} = -\nabla V(x) - \gamma \dot{x} + \sigma \dot{B}. \quad (4.53)$$

The coordinate x in this equation stands for cartesian coordinates of the four extended atoms making up the butane molecule, see [17] for details of the CHARMM forcefield used here.

From a chemical point of view interest is focused on the dihedral angle, which is the angle between the two planes in \mathbb{R}^3 formed by atoms 1,2,3 and atoms 2,3,4 respectively; see the sketch in figure 4.10. Conformational change is manifest in this

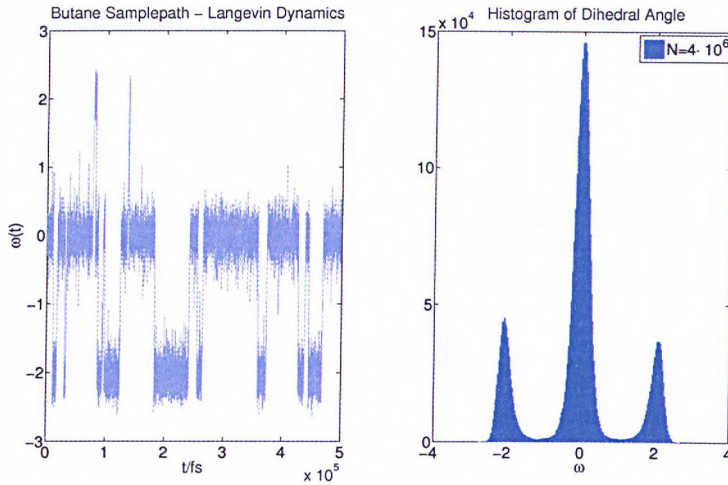


Figure 4.11: MD Samplepath: Butane

angle, and the cartesian coordinates themselves are of little direct chemical interest. Hence it is natural to try and describe the stochastic dynamics of the dihedral angle in a self-contained fashion.

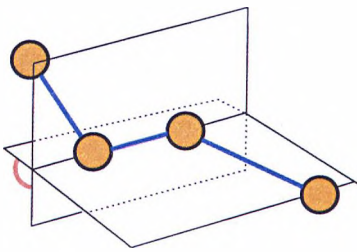


Figure 4.10: Sketch of Dihedral Angle

One MD run is produced using a timestep of $\Delta t = 10^{-16}$ s (one tenth of a femtosecond) and a Verlet variant (see p.435 in [54]) covering a total time of $T = 4 \cdot 10^{-9}$ s (4 nanoseconds). A section of path of the dihedral angle versus time can be seen on the left of figure 4.11; the corresponding histogram is depicted to the right of that figure. It is known ([23]) that the stationary distribution of (4.53) is given by the canonical distribution associated with the torsional potential, so that an explicit analytical representation can easily

be obtained.

It should be stressed that the effective stochastic differential equation governing the behaviour of the dihedral angle ω is *not* of the form (4.15), in particular, it will have a non-constant diffusivity σ . So, fitting to this data tests the robustness of the fitting

algorithm in a way that the experiments in previous sections did not.

4.8.2 Fitting

The physical time-units in seconds are miniscule and do not lead to SDE parameter fits of order one. It transpires that, in order to obtain parameter values of order one, rescaling time so that the final time becomes $T = 80000$ is a good choice. This rescaling is useful in comparing convergence properties with what was observed in section 6.

In order to assess consistency, the MD data is subsampled, at timesteps $\Delta t \in \{1 \cdot 10^{-15}\text{s}, 2 \cdot 10^{-15}\text{s}, 3 \cdot 10^{-15}\text{s} \dots\}$ in physical time units, corresponding to $\{k \cdot 0.02\}_{k \in \mathbb{N}}$ in the rescaled time units. The Deterministic Scan Gibbs sampler is then run for $N_{\text{Gibbs}} = 40$ outer iterations on each path using a potential ansatz

$$V(\omega) = \sum_{k=1}^c \theta_k \cos^k(\omega)$$

where $c \in \{3, 5, 7\}$ is used. This corresponds to a choice of the force function in (4.15). The obtained drift parameters under subsampling at timestep Δt can be seen from figure 4.12. This plot shows the behaviour of the drift parameters averaged over $N_{\text{Gibbs}} = 100$ Monte-Carlo samples $\theta_1, \dots, \theta_5, \gamma$ as the subsampling rate is increased. Below a subsampling rate $k = 20$, behaviour consistent with $\mathcal{O}(\Delta t)$ errors is observed indicating convergence of the algorithm as Δt is decreased. This is exactly the behaviour observed on simulated data and it is a measure of the robustness of the proposed algorithm.

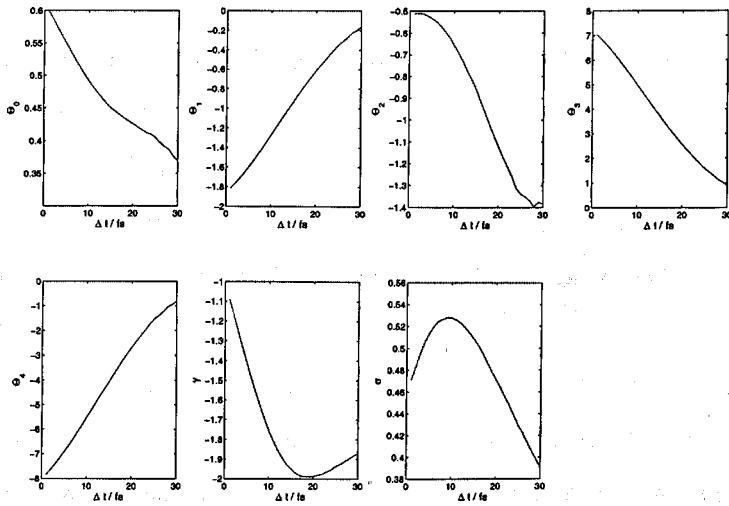


Figure 4.12: Convergence for fitted MD path with subsampling

4.8.3 Limitations

The desirable convergence properties of the algorithm in Δt and T should not be confused with inference about whether fitting this kind of model to this kind of MD data gives a good or a bad fit, it merely indicates that, using the algorithm suggested in this chapter, it is possible to perform such fitting.

To show limitations of the model in this particular application and see how the performance can be assessed using the fitting algorithm from section 4.7.2, we show posterior invariant probability densities resulting from the fitted trigonometric potentials. In order to do this, we convert the drift parameter samples $\{D_j^{(m)}\}_{j=1}^c$ obtained at step m using input data subsampled at rate $k = 1$ to an invariant density, $\rho^{(m)}$ specified by its values on an equidistant grid on the interval $[-\pi, \pi]$. These densities for $m \in \{1, \dots, 1000\}$ are then averaged and their standard deviation is computed pointwise on the grid. This results in the plot given in figure 4.13. There, we display results for three orders of trigonometric potential c to be fitted and contrast this with the empirically observed invariant density and the density arising from the classic canonical thermodynamic ensemble which is proportional to $\exp\left(-\frac{V(\omega)}{kT}\right)$. For the pa-

parameterisation used here, it is known that the latter two agree in the limit $T \rightarrow \infty$, see [23].

With increasing polynomial order c we find some qualitative change in the resulting probability and also (in particular moving from $c = 5$ to $c = 7$) a marked increase in posterior variance. This goes hand-in-hand with a marked increase in the condition number of the drift parameter matrix M_E in (4.36). It is simply an ill-conditioned problem to derive higher and higher order polynomial coefficients from a fixed length of observed path.

It is observed that even though the empirically observed invariant density is smooth and close to the thermodynamical expectation, the fitted potentials induce an SDE whose invariant measure is not a good approximation of the empirical density. This may simply be attributed to the fact that the SDE that is being fitted does not represent a good model of the *dynamics* of the dihedral angle in the butane molecule with second order Langevin heat bath model. One crucial qualitative difference in the dynamics is the fact that the butane molecule is described by a (high dimensional) SDE with *multiplicative* noise whereas an additive noise model is being fitted. This will be further elucidated in section 5.8.6.

4.9 Conclusions

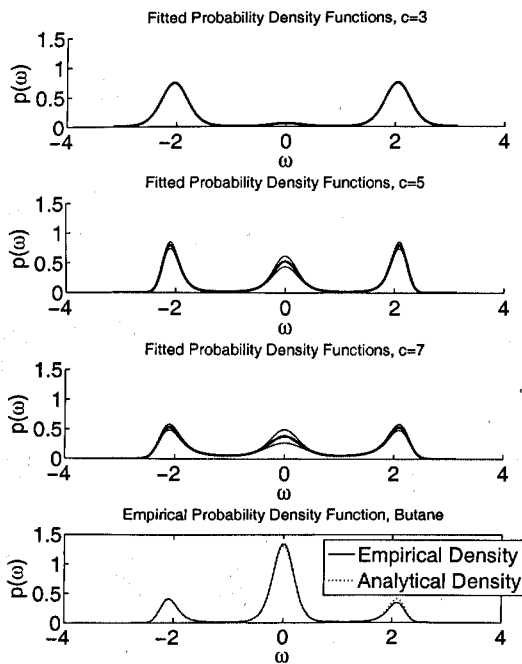


Figure 4.13: PDFs resulting from fitted potentials for different orders of trigonometric potential - Shaded regions display posterior variance

pass order k hypoelliptic problems and it has been tested successfully on a third order example. Furthermore, non-linear p -dependence in the example (4.11) can be dealt with using a Langevin sampler for the missing path and this has also been tested.

Further avenues of investigation include the use of imputed data-points between samples to diminish $\mathcal{O}(\Delta t)$ errors; however there is a risk of bad mixing as σ is determined by the small scale behaviour of the process which would then be dominated by the imputed data points. This has been analysed in the case of elliptic diffusion processes in [53].

A hybrid algorithm for fitting drift and diffusion parameters of a hypoelliptic diffusion process with constant diffusivity from observation of smooth data at discrete times has been described. Its performance has been validated numerically for a number of test cases and an application to molecular dynamics data has been given. While parameter fitting can be viewed as an inverse problem for SDE solvers – and thus ill-conditioning of some kind is always to be expected – a detailed understanding of the ill-conditioning induced by hypoellipticity and partial observation has been attained.

While only second order hypoelliptic problems have been treated in this article, the algorithm's applicability is expected to encom-

Chapter 5

Nonparametric Estimation for Diffusion Processes

5.1 Overview

In applications such as molecular dynamics it is of interest to fit Langevin-like equations to data. Practitioners do this by a variety of *ad hoc* procedures such as fitting to the empirical measure generated by the data, and fitting to properties of auto-correlation functions. Statisticians, on the other hand, have well-developed estimation procedures which fit diffusion processes to data applying the maximum likelihood principle to the path-space density of the desired model equations, and through knowledge of the properties of the quadratic variation. In this chapter we show that the procedures used by practitioners and statisticians are, in fact, closely related. We do this by introducing a nonparametric approach to estimation for diffusion processes. Furthermore, we present the results of numerical experiments which probe the relative efficacy of the two approaches to model identification.

5.2 Introduction

In many applications beyond molecular dynamics (econometrics, atmospheric sciences, signal processing) it is of interest to fit a diffusion process to a time-series. The purpose

of this chapter is to introduce a non-parametric approach to this estimation procedure. We will focus here on reversible processes and non-reversible processes of second order Langevin-type. Thus, applications to molecular dynamics are of particular relevance.

The basic idea behind nonparametric drift estimation is to express the pathspace likelihood for the diffusion process in terms of integrals across the state space of the diffusion, rather than the usual time integrals. In the state space integrals, the information about the time-series appears through the empirical density that it generates. Applying standard calculus of variation techniques to maximise these expressions for the likelihood then leads to non-parametric estimation of the drift, with estimates given in terms of the empirical density.

We will show that this approach leads to methods closely related to a variety of estimation procedures appearing in the literature, in particular to the minimum distance estimator (MDE) and to techniques commonly used by practitioners in molecular dynamics based around fitting to the empirical invariant measure.

Whilst it is statistical folklore that drift estimation is considerably harder than diffusion estimation (see e.g. [50], [62]), in that the quadratic variation in principle reveals the diffusion coefficient, it is common practical experience with real data that diffusion estimation is the harder part of the problem (see e.g. [34], [59]). In this context, we discuss a variety of different approaches to the estimation of the diffusion co-efficient, comparing standard statistical procedures and those used by practitioners.

Our setting is to work with diffusions of the form

$$\frac{dx}{dt} = b(x) + \sqrt{2K(x)} \frac{dW}{dt} \quad (5.1)$$

where W is a standard d -dimensional Brownian Motion, x is a stochastic process adapted to the Brownian Motion, $K : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is a symmetric positive-semidefinite valued function, and $b : \mathbb{R}^d \rightarrow \mathbb{R}^d$. We assume ergodicity of the stochastic process x .

The chapter is organised as follows. In section 3 we describe non-parametric drift estimation for gradient diffusions where results for finite observation times can be attained. In section 4 we generalise to drift estimation for reversible processes stating results in the limit of long observation times, and section 5 contains a similar development for second-order Langevin equations, an important class of non-reversible processes.

Section 6 discusses various methods for estimating the diffusion coefficient. In section 7 we comment on the relationship between the material in sections 3-5, and the existing literature. Section 8 contains numerical experiments in which we compare the efficacy of the nonparametric estimation procedures derived here with standard statistical procedures.

We conclude the section by discussing some properties of the diffusion (5.1) which are pertinent in what follows. In so doing we describe the basic idea underlying non-parametric drift estimation.

Given an invertible covariance matrix $R \in \mathbb{R}^{d \times d}$ we define an inner product and norm on \mathbb{R}^d by

$$\begin{aligned}\langle a, b \rangle_R &= a^T R^{-1} b \quad \forall a, b \in \mathbb{R}^d, \\ |a|_R^2 &= \langle a, a \rangle_R \quad a \in \mathbb{R}^d.\end{aligned}$$

Let z solve equation (5.1) with $b = 0$ so that

$$\frac{dz}{dt} = \sqrt{2K(z)} \frac{dW}{dt} \quad (5.2)$$

and let \mathbb{P} and \mathbb{Q} be the pathspace measures generated by (5.1) and (5.2) on $[0, t]$. Then these measures are absolutely continuous with Radon-Nikodym derivative

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp(-TI(b)) \quad (5.3)$$

where

$$I(b) = -\frac{1}{4T} \int_0^T \left(|b(x)|_{K(x)}^2 dt - 2\langle b(x), dx \rangle_{K(x)} \right) \quad (5.4)$$

Recall that the *generator* for the process (5.1) is the operator

$$\mathcal{L} := b \cdot \nabla + K : \nabla \nabla, \quad (5.5)$$

and that $v(x, t) = \mathbb{E}\varphi(x(t)|x(0) = x)$ solves

$$\begin{aligned}\frac{\partial v}{\partial t} &= \mathcal{L}v, & t > 0, \\ v &= \varphi, & t = 0.\end{aligned} \quad (5.6)$$

Probability densities $\varrho(x, t)$ for x solving (5.1) satisfy (see [9], [51]) the Fokker-Planck equation

$$\begin{aligned}\frac{\partial \varrho}{\partial t} &= \mathcal{L}^* \varrho, & t > 0, \\ \varrho &= \varrho_0, & t = 0,\end{aligned}\tag{5.7}$$

where ϱ_0 is the initial density for $x(0)$.

By ergodicity we know that

$$\lim_{t \rightarrow \infty} I(b) = -\frac{1}{4} \int_{\mathbb{R}^d} \left(\varrho(x) |b(x)|_{K(x)}^2 - 2 \langle b(x), \varrho(x) \mathcal{L}x \rangle_{K(x)} \right) dx$$

Of course, we do not know $\varrho(x)$ and $\varrho(x) \mathcal{L}x$ exactly – we only have the time series $\{x(s)\}_{s \in [0, t]}$. However, we can approximate $\varrho(x)$ by the empirical density $\hat{\varrho}(x)$ generated by this time series. If we can also approximate $\varrho(x) \mathcal{L}x$ in terms of the data, say by an expression $r(x)$, then we approximate $I(b)$ by

$$I(b) \approx I_a(b) := -\frac{1}{4} \int_{\mathbb{R}^d} \left(\hat{\varrho}(x) |b(x)|_{K(x)}^2 - 2 \langle b(x), r(x) \rangle_{K(x)} \right) dx.\tag{5.8}$$

Maximising $I_a(b)$ then gives a non-parametric estimate of $b(x)$, say $\hat{b}(x)$. Since $I_a(\cdot)$ is a quadratic functional, the optimisation problem can be solved explicitly as follows. We have

$$I_a(b + \delta b) = I_a(b) - \frac{1}{4} \int_{\mathbb{R}^d} \hat{\varrho}(x) |\delta b|_{K(x)}^2 dx + \frac{1}{2} \int_{\mathbb{R}^d} \langle r(x) - \hat{\varrho}(x) b(x), \delta b(x) \rangle_{K(x)} dx.\tag{5.9}$$

From this expression it is clear that $I_a(b)$ is maximised by choosing $b(x) = \hat{b}(x)$ to be

$$\hat{b}(x) = \frac{1}{\hat{\varrho}(x)} r(x).\tag{5.10}$$

Our ability to carry out this program depends upon our ability to approximate $\varrho(x) \mathcal{L}x$ by $r(x)$ given only knowledge of the time series. We discuss this issue in sections 3 and 4, motivated by the examples in section 2.

We conclude this section with a few remarks on the Fokker-Planck equation (5.7) for (5.1), and relatedly on reversibility. This equation may be written in the form

$$\begin{aligned}\frac{\partial \varrho}{\partial t} &= \nabla \cdot (l(\varrho)), \\ l(\varrho) &= -b\varrho + \nabla \cdot (K\varrho).\end{aligned}\tag{5.11}$$

The quantity $l(\varrho)$ is known as the probability current ([9], eq. (5.2.8), p.119). The steady solution $\varrho(x)$ satisfies

$$\begin{aligned} 0 &= \nabla \cdot (l(\varrho)), \\ l(\varrho) &= -b\varrho + \nabla \cdot (K\varrho). \end{aligned} \tag{5.12}$$

The process is *reversible* if the steady solution $\varrho(x)$ is in the null-space of l so that the probability current is zero: $l(\varrho(x)) = 0$.

5.3 The Gradient Case

One of the motivational examples which we will use to illustrate our work is a gradient diffusion of the form

$$\frac{dx}{dt} = -\nabla V(x) + \sqrt{2k} \frac{dW}{dt}, \tag{5.13}$$

where V is a sufficiently smooth and confining potential and $k \in \mathbb{R}^+$ is a constant. Note that (5.13) is a special case of (5.1). Also, it should be highlighted that (5.13) is of the type studied parametrically in chapter 3, albeit potentially in higher dimension. In the case of this example, and of gradient diffusion processes in general, a direct link can be made between the maximum likelihood estimator and the practitioners' way of fitting the empirical density. To do this, we use the Radon-Nikodym derivative (5.3) and convert the Ito integral in (5.4) to a Stratonovich integral. Using $\nabla V = b$ we obtain

$$I(x) = \frac{1}{T} \int_0^T \langle \nabla V(x), \circ dx \rangle + \frac{1}{2T} \int_0^T \left(|\nabla V(x)|^2 - 2k \Delta V(x) \right) dt. \tag{5.14}$$

This will be pursued in detail in this section and results that hold even for finite times of observation T will be given in the 1D case which corresponds exactly to the processes treated in chapter 3. We will generalise to reversible processes using a different argument in the next section.

5.3.1 Statisticians' Approach

The maximum likelihood approach to this would be to write $\nabla V(x)$ is a linear combination of basis functions $f_i(x)$, so that

$$\nabla V(x) = \sum_{i=1}^c \theta_i f_i(x).$$

Then to substitute this into the expression (5.4) and maximise with respect to θ . Since $I(x)$ is quadratic in θ in this case, this gives rise to a system of linear equations.

5.3.2 Practitioners' Approach

The invariant measure for (5.13) is proportional to $\exp(-\frac{1}{k}V(x))$. The typical approach of the practitioner is to fit $V(x)$ to the logarithm of the empirical measure generated by the path $\{x(t)\}_{t \in [0, T]}$. This appears very different from what a statistician would do, but is in fact closely related. To see this we attempt a non-parametric estimation of the drift potential $V(x)$ via the maximum likelihood principle, based on minimising $I(x)$ given by (5.14). Now

$$\begin{aligned} I(x) &= \frac{1}{T} \int_0^T \langle \nabla V(x), \circ dx \rangle + \frac{1}{2T} \int_0^T (|\nabla V(x)|^2 - 2k\Delta V(x)) dt \\ &= \frac{1}{T} (V(x(T)) - V(x(0))) + \frac{1}{2T} \int_0^T (|\nabla V(x)|^2 - 2k\Delta V(x)) dt. \end{aligned} \quad (5.15)$$

Under suitable assumptions on the potential the first term tends to zero almost surely as $T \rightarrow \infty$. Thus for large T it is natural to estimate $V(x)$ by minimising

$$\frac{1}{2T} \int_0^T (|\nabla V(x)|^2 - 2k\Delta V(x)) dt.$$

For large T we approximate the time-average by average against the empirical measure with density $\hat{\rho}$. This suggests that we minimise the following functional of $V(x)$, namely

$$\mathcal{I}(V) = \frac{1}{2} \int_{\mathbb{R}^d} (|\nabla V(x)|^2 - 2k\Delta V(x)) \hat{\rho}(x) dx.$$

Now

$$\mathcal{I}(V + W) = \mathcal{I}(V) + \int_{\mathbb{R}^d} \langle \hat{\rho}(x) \nabla V(x) + k \nabla \hat{\rho}(x), \nabla W \rangle dx + \mathcal{I}(W).$$

Hence $\mathcal{I}(V)$ is minimised where

$$\hat{\rho}(x) \nabla \hat{V}(x) + k \nabla \hat{\rho}(x) = 0.$$

Assuming that the empirical measure is zero at infinity we see that

$$\hat{V}(x) = -k \log \hat{\rho}(x). \quad (5.16)$$

A practitioners' approach would involve fitting $V(x)$ to the logarithm of the empirical density and hence can be thought of as invoking a maximum likelihood principle, as in the statisticians' approach. Note, however, that this fit determines $V(x)$ only up to a time constant, k . We show how to estimate this time-constant in section 5.6 below.

5.3.3 Specialising to 1D - finite T results

In the one-dimensional case, rather than moving to the limit of long observations, $T \rightarrow \infty$, it can be instructive to employ the local time L_t^a of the process (5.13), which intuitively corresponds to the time spent at a up to time t .

Theorem 2.11.7 in [13] states that for x being a 1-d continuous semimartingale with local time L_t^a the following identity holds for any Borel-measurable, bounded function g :

$$\int_{-\infty}^{\infty} L_t^a g(a) da = \int_0^t g(x_s) d\langle x \rangle_s \quad (5.17)$$

Note that for the process (5.13) we have

$$dt = \frac{1}{2k} d\langle x \rangle_t$$

so that the new integral becomes

$$I(x) = \frac{1}{T} (V(x(T)) - V(x(0))) + \frac{1}{2T} \int_{\mathbb{R}} \left(\frac{1}{2k} |V'(a)|^2 - V''(a) \right) L_t^a da,$$

where first and second derivatives of the potential V have been expressed as $V'(x) = \frac{d}{dx} V(x)$ and $V'' = \frac{d^2}{dx^2} V(x)$. It can be shown that the local time L_t^a is jointly continuous in (t, a) , however it is not in general differentiable. By looking at a suitable weak interpretation of the sequel, it might be possible to disregard this technical problem. So we integrate by parts to obtain:

$$I(V) = \frac{1}{T} (V(x(T)) - V(x(0))) + \frac{1}{2T} \int_{\mathbb{R}} \frac{1}{2k} |V'(a)|^2 L_t^a + V'(a) \cdot \frac{d}{da} L_t^a da \quad (5.18)$$

Consider the variational derivative of (5.18):

$$\begin{aligned} I(V + \varepsilon W) = I(V) + \frac{\varepsilon}{T} (W(x(T)) - W(x(0))) + \frac{1}{2T} \int_{\mathbb{R}} \left(\varepsilon \frac{1}{k} V'(a) \cdot W'(a) L_t^a \right. \\ \left. + \frac{\varepsilon}{T} W'(a) \cdot \frac{d}{da} L_t^a \right) da + \mathcal{O}(\varepsilon^2) \end{aligned}$$

Rewriting the point values $W(x(t))$ using Dirac- δ -distributions and integrating these partially results in the following expression:

$$T \frac{\delta I}{\delta V}(W) = \int_{\mathbb{R}} \left(-W'(a)H_{x(T)}(a) + W'(a)H_{x(0)}(a) + \frac{1}{2k}V'(a) \cdot W'(a)L_t^a + \frac{1}{2}W'(a) \cdot \frac{d}{da}L_t^a \right) da$$

So the non-parametric estimator of the gradient of the potential V is given by:

$$V' = 2k \frac{d}{da} \log L_t^a + \frac{H_{x(0)} - H_{x(T)}}{L_t^a} \quad (5.19)$$

Given that the Heaviside function will be zero in the far negative, this could be integrated. This equation should be compared to (5.16).

Since this functional is also used to characterise the maximum likelihood estimator for parametric inference for V , this shows a close link between the two estimators. For finite final time T this link is perturbed by the Heaviside functions in (5.19). This perturbation, however, is typically of order $\mathcal{O}(T^{-1})$ whereas the average (root mean square) deviation of the first term in (5.19) is $\mathcal{O}(\sqrt{T^{-1}})$, so that this link for finite final time reveals a much closer relationship than the limiting arguments above would lead one to believe.

5.3.4 Extension to higher dimensions?

Extending local time to multi-D is not feasible via the trick using the Meyer-Tanaka formula. One could attempt to define a random measure on \mathbb{R}^d that still fulfils (5.17), however this would not normally be continuous in space, so taking its gradient and integrating V against it might present a technical problem.

5.4 Drift Estimation for Reversible Processes

We describe nonparametric estimation of $b(x)$ in (5.1) assuming that $K(x)$ is known, and that the process is reversible. Notice that $\mathcal{L}x$, with generator \mathcal{L} given by (5.5), is given by $\mathcal{L}x = b(x)$. Thus

$$\varrho(x)\mathcal{L}x = \varrho(x)b(x) \quad (5.20)$$

Recall that we wish to approximate $\varrho(x)\mathcal{L}x$ in terms of the time series data alone. The identity (5.20) fails to do this because $b(x)$ is not known to us – we wish to estimate it. However, if the process is reversible then, from (5.12),

$$b(x)\varrho(x) = \nabla \cdot (K(x)\varrho(x)) \quad (5.21)$$

and so we have

$$\varrho(x)\mathcal{L}x = \nabla \cdot (K(x)\varrho(x)). \quad (5.22)$$

Since $K(x)$ is assumed to be known we deduce that we may approximate $\varrho(x)\mathcal{L}x$ by $r(x) = \nabla \cdot (K(x)\hat{\varrho}(x))$. The approximate scaled log-likelihood given by (5.8) is a quadratic functional of $b(x)$. Thus, using formula (5.9) we obtain for \hat{b} maximising $I_a(\cdot)$ from:

$$\hat{b}(x) = \frac{1}{\hat{\varrho}(x)}r(x) = \frac{1}{\hat{\varrho}(x)}\nabla \cdot (K(x)\hat{\varrho}(x)) \quad (5.23)$$

The identity (5.23) provides our non-parametric estimate of $b(x)$. We make several remarks.

1. The expression (5.23) shows that, provided $\hat{\varrho}(X) \rightarrow \varrho(x)$ as $T \rightarrow \infty$ in an appropriate function space including derivatives, then $\hat{b}(x) \rightarrow b(x)$, by (5.21).
2. To ensure convergence of $\hat{\varrho}(x)$ to $\varrho(x)$, and in particular convergence of derivatives is, in general, non-trivial.
3. For the case $K(x) = K \in \mathbb{R}^{d \times d}$, a symmetric positive-definite matrix, independent of x , equation (5.21) shows that for a reversible process

$$b(x) = -K\nabla V(x) \quad (5.24)$$

for some scalar potential $V : \mathbb{R}^d \rightarrow \mathbb{R}$. In this case, too, the non-parametric estimate (5.23) can be written as

$$\hat{b}(x) = K\nabla \log \hat{\varrho}(x). \quad (5.25)$$

Since we know that the true drift has the form given by (5.24), it is natural to estimate $\hat{b}(x)$ non-parametrically as $\hat{b}(x) = -K\nabla \hat{V}(x)$. Then (5.25) implies that

$$\hat{V}(x) = -\log \hat{\varrho}(x). \quad (5.26)$$

4. It is reassuring to find that (5.26) is the same as (5.16) so that our argument for reversible processes generalises the argument put forward in the gradient case.

5.5 Drift Estimation for Second Order Langevin Equations

5.5.1 Direct Variational Approach

We now consider an example of a non-reversible process: the second order Langevin equation

$$\frac{d^2q}{dt^2} + k\beta\frac{dq}{dt} + \nabla V(q) = \sqrt{2k}\frac{dW}{dt}, \quad (5.27)$$

where β is the damping constant, k is the diffusivity and W is standard Brownian motion. Note that (5.27) is a special case of (5.1). Also, it should be highlighted that (5.27) is the process studied parametrically in chapter 4. Mimicking the presentation in the first order case, we will first present a direct approach applying variational calculus to the Radon-Nikodym derivative in the gradient case and then move to the more general framework using the Kolmogorov equation (5.6).

If we set $p = \frac{dq}{dt}$ then from (5.27) we obtain the following system of equations:

$$\begin{aligned} \frac{dq}{dt} &= p, \\ \frac{dp}{dt} &= -k\beta p - \nabla V(q) + \sqrt{2k}\frac{dW}{dt}. \end{aligned} \quad (5.28)$$

The Radon-Nikodym derivative of the measure on path-space for (4.1) with respect to the measure generated by

$$\frac{dp}{dt} = \sqrt{2k}\frac{dW}{dt} \quad (5.29)$$

is proportional to

$$\exp\left(-\frac{1}{2k}I(q, p)\right) \quad (5.30)$$

where

$$I(q, p) = \int_0^T \langle k\beta p + \nabla V(q), dp \rangle + \frac{1}{2} \int_0^T (|\nabla V(q)|^2 + (k\beta)^2 |p|^2 + 2k\beta \langle \nabla V(q), p \rangle) dt. \quad (5.31)$$

Statisticians' Approach

As in the first order case, a statisticians' maximum likelihood approach would be to represent

$$\nabla V(q) = \sum_{i=1}^N \theta_i f(q)$$

and then to minimise $I(q, p)$ with respect to θ .

Practitioners' Approach

The invariant measure for (5.28) is a product measure, Gaussian $\mathcal{N}(0, \frac{1}{\beta}I)$ in p and identical to that arising in the first order case in q , namely proportional to $\exp(-\beta V(q))$:

$$\begin{aligned} \rho(q, p) &= C \exp(-\beta V(q)) \exp\left(-\frac{1}{2\beta} \|p\|_2^2\right) \\ &= \rho(q) g(p) \end{aligned}$$

As in the first order case we attempt a non-parametric estimation of the drift potential $V(q)$ via the maximum likelihood principle. This suggests minimising $I(q, p)$. An integration by parts shows that

$$\begin{aligned} \frac{1}{T} I(q, p) &= \frac{1}{T} \int_0^T \langle k\beta p + \nabla V(q), dp \rangle \\ &\quad + \frac{1}{2T} \int_0^T (|\nabla V(q)|^2 + (k\beta)^2 |p|^2 + 2k\beta \langle \nabla V(q), p \rangle) dt \\ &= -\frac{1}{T} \int_0^T D^2 V(q) : p \otimes p dt \\ &\quad + \frac{1}{2T} \int_0^T (|\nabla V(q)|^2 + (k\beta)^2 |p|^2 + 2k\beta \langle \nabla V(q), p \rangle) dt + \frac{1}{T} \int_0^T k\beta \langle p, dp \rangle. \end{aligned}$$

For large T we approximate the time-averages in the non-stochastic integrals by average against the empirical measure with density $\hat{\rho}(q)\hat{g}(p)$, assuming some algorithm is used to approximately factorise the empirical measure. This suggests that we minimise the following functional of $V(q)$:

$$\begin{aligned} \mathcal{I}(V) &= - \int_{\mathbb{R}^{2d}} (D^2 V(q) : p \otimes p) \hat{\rho}(q) \hat{g}(p) dq dp + \frac{1}{2} \int_{\mathbb{R}^d} (|\nabla V(q)|^2) \hat{\rho}(q) dq \\ &\quad + \frac{1}{2} \int_{\mathbb{R}^d} (k\beta)^2 |p|^2 \hat{g}(p) dp + \int_{\mathbb{R}^{2d}} k\beta \langle \nabla V(q), p \rangle \hat{\rho}(q) \hat{g}(p) dq dp + \frac{1}{T} \int_0^T k\beta \langle p, dp \rangle. \end{aligned}$$

The fourth integral is zero if we impose the condition that $\mathbb{E}p = 0$ and the first integral simplifies to an integral over q alone if we impose $\mathbb{E}p \otimes p = \frac{1}{\beta}I$. This suggests that we minimise

$$\mathcal{I}(V) = \frac{1}{2} \int_{\mathbb{R}^d} (|\nabla V(q)|^2 - \frac{2}{\beta} \Delta V(q)) \hat{\rho}(q) dq + \frac{1}{2} (k\beta)^2 \int_{\mathbb{R}^d} |p|^2 \hat{g}(p) dp + \frac{1}{T} \int_0^T k\beta \langle p, dp \rangle.$$

Setting the variation of $\mathcal{I}(V)$ with respect to $V(q)$ to zero, we obtain

$$\hat{V}(q) = -\frac{1}{\beta} \log \hat{\rho}(q)$$

as in the first order case. Thus, the non-parametric maximum likelihood principle for estimation of $V(q)$ leads to the fitting of $V(x)$ to the empirical measure in x . The damping parameter β can also be estimated in this setting.

5.5.2 Langevin in the general Framework

We now show how to integrate the previous subsection into the general framework of (5.1) as well as how to incorporate estimation of the damping. Firstly, the process (5.27) can be cast in the general framework of this chapter as follows:

Let $x \in \mathbb{R}^{2d}$ with

$$b(x) = -\beta K(x) \nabla H(x) + J \nabla H(x) \quad (5.32)$$

where

$$x = \begin{pmatrix} q \\ p \end{pmatrix}, \quad H(x) = \frac{1}{2} p^2 + V(q), \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then

$$\nabla H(x) = \begin{pmatrix} \nabla V(q) \\ p \end{pmatrix}, \quad J \nabla H(x) = \begin{pmatrix} p \\ -\nabla V(q) \end{pmatrix}. \quad (5.33)$$

Assume

$$(\nabla \cdot K)(x) = 0. \quad (5.34)$$

The stationary measure for this process is then given by $\varrho(x) = C \exp(-\beta H(x))$. To

see this, note that

$$\nabla \varrho(x) = -\beta \nabla H(x) \varrho(x), \quad (5.35)$$

$$-b(x) \varrho(x) = -K(x) \nabla \varrho(x) + \frac{J}{\beta} \nabla \varrho(x) \quad (5.36)$$

$$= -\nabla \cdot (K(x) \varrho(x)) + \frac{J}{\beta} \nabla \varrho(x). \quad (5.37)$$

Thus $l(\varrho(x)) = \frac{1}{\beta} J \nabla \varrho(x)$ and, since skew-gradients are divergence free,

$$\nabla \cdot l(\varrho(x)) = 0. \quad (5.38)$$

Equation (5.38) shows that $\varrho(x)$ is stationary by (5.11). Note that since $l(\varrho(x))$ is not identically zero, the process is not reversible.

Now we ask whether we can express $r(x) \mathcal{L}x = \varrho(x) b(x)$ in (5.8) purely in terms of time-series data in this case. From (5.36) we deduce that

$$-b(x) \varrho(x) = -K(x) \nabla \varrho(x) + \left(\int p^2 \varrho(x) dx \right) J \nabla \varrho(x),$$

since p is Gaussian with distribution $\mathcal{N}(0, \frac{1}{\beta})$ under the stationary measure. This suggests that we approximate $\varrho(x) b(x)$ by

$$r(x) = K(x) \nabla \hat{\varrho}(x) - \left(\int p^2 \hat{\varrho}(x) dx \right) J \nabla \hat{\varrho}(x). \quad (5.39)$$

We may use this expression in (5.10) to estimate $b(x)$ nonparametrically.

5.5.3 Nonparametric Estimation of $b(x)$

With the above definition of $r(x)$, we deduce from (5.10) that b is maximised where $b(x) = \hat{b}(x)$:

$$\hat{\varrho}(x) \hat{b}(x) = K(x) \nabla \hat{\varrho}(x) - \left(\int p^2 \hat{\varrho}(x) dx \right) J \nabla \hat{\varrho}(x) \quad (5.40)$$

The identity (5.40) provides our non-parametric estimate of $b(x)$. We make several remarks.

1. If $\nabla \hat{\varrho}(x) \rightarrow \nabla \varrho(x)$ and $\hat{\varrho}(x) \rightarrow \varrho(x)$ as $t \rightarrow \infty$ then $\nabla \hat{\varrho}(x) \rightarrow -\beta \nabla H(x) \varrho(x)$, and $\int p^2 \hat{\varrho}(x) dx \rightarrow \frac{1}{\beta}$. Hence $\hat{b}(x) \rightarrow b(x)$ by (5.40).

2. As in the reversible case, it is of interest to estimate the potential $V(q)$ non-parametrically. Since $V(q)$ and β together determine $b(x)$ given by (5.40) it is in fact natural to estimate $(\beta, V(q))$; we study this question in the next subsection.
3. The specific instance of the second order Langevin equation (5.27) corresponds to a singular diffusion matrix K . However, the next subsection will show that this singularity can be handled in the general context of this section, recovering the calculations of section 2.

5.5.4 Nonparametric Estimation of $(V(q), \beta)$

We study the situation above in the case where

$$K(x) = \begin{pmatrix} K_1(q) & 0 \\ 0 & K_2(q) \end{pmatrix}$$

and we introduce the notation

$$r(x) = \begin{pmatrix} r_1(x) \\ r_2(x) \end{pmatrix}, \quad \tilde{\beta} = \left(\int p^2 \hat{\rho}(x) dx \right)^{-1}.$$

We start with the assumption that $K(x) = K(q)$ is positive definite uniformly on \mathbb{R}^d . However, we will show that, when estimating $(V(q), \beta)$ only, the singular limit of $K_1(q) \rightarrow 0$ may be taken. Rather than trying to estimate b we try to estimate $(\beta, V(q))$; together these quantities determine $b(x)$. Recall the functional we wish to minimise, $I_a(b)$ from (5.8). To understand how I_a depends on (β, V) we calculate the two terms under the integral in (5.8). Firstly, we have

$$\begin{aligned} |b(x)|_{K(x)}^2 &= \beta^2 |\nabla H(x)|_{K(x)^{-1}}^2 + |J\nabla H(x)|_{K(x)}^2 \\ &= \beta^2 |\nabla V(q)|_{K_1(q)^{-1}}^2 + \beta^2 |p|_{K_2(q)^{-1}}^2 + |p|_{K_1(q)}^2 + |\nabla V(q)|_{K_2(q)}^2 \end{aligned} \quad (5.41)$$

Also

$$\begin{aligned} \langle b(x), r(x) \rangle_{K(q)} &= \langle -\beta K(x) \nabla H(x) + J \nabla H(x), r(x) \rangle_{K(q)} \\ &= -\beta \langle \nabla V(q), r_1(x) \rangle - \beta \langle p, r_2(x) \rangle + \langle p, r_1(x) \rangle_{K_1(q)} \\ &\quad - \langle \nabla V(q), r_2(x) \rangle_{K_2(q)}. \end{aligned} \quad (5.42)$$

In the singular limit $K_1(q) = 0$ there are two terms in the preceding expressions for $|b(x)|_{K(q)}^2$ and $\langle b(x), r(x) \rangle_{K(q)}$ which become unbounded. However, neither depend upon β and $V(q)$; they are hence irrelevant to the likelihood calculation and we ignore them. Further simplifications to $I(b)$ are possible, using the structure of the invariant measure. Notice that

$$\begin{aligned}\int_{\mathbb{R}^{2d}} A(q)p\varrho(x)dx &= 0 \\ \int_{\mathbb{R}^{2d}} \langle a(q), \nabla_p \varrho(x) \rangle dx &= 0 \\ \int_{\mathbb{R}^{2d}} \langle p, \nabla_p \varrho(x) \rangle_{K_2(q)^{-1}} dx &= -\beta \int_{\mathbb{R}^{2d}} \varrho(x)|p|_{K_2(q)^{-1}}^2 dx.\end{aligned}$$

Thus in the following we will make the substitutions

$$\begin{aligned}\int_{\mathbb{R}^{2d}} A(q)p\hat{\varrho}(x)dx &\mapsto 0 \\ \int_{\mathbb{R}^{2d}} \langle a(q), \nabla_p \hat{\varrho}(x) \rangle dx &\mapsto 0 \\ \int_{\mathbb{R}^{2d}} \langle p, \nabla_p \hat{\varrho}(x) \rangle_{K_2(q)^{-1}} dx &\mapsto -\tilde{\beta} \int_{\mathbb{R}^{2d}} \hat{\varrho}(x)|p|_{K_2(q)^{-1}}^2 dx.\end{aligned}\tag{5.43}$$

Now, from (5.39),

$$\begin{aligned}r_1 &= K_1(q)\nabla_q \hat{\varrho}(x) - \frac{1}{\tilde{\beta}}\nabla_p \hat{\varrho}(x), \\ r_2 &= K_2(q)\nabla_p \hat{\varrho}(x) + \frac{1}{\tilde{\beta}}\nabla_q \hat{\varrho}(x).\end{aligned}$$

Hence, applying (5.43) to (5.42), we obtain

$$\begin{aligned}\int_{\mathbb{R}^{2d}} \langle \nabla V(q), r_1(x) \rangle dx &\mapsto \int_{\mathbb{R}^{2d}} \langle \nabla V(q), \nabla_q \hat{\varrho}(x) \rangle_{K_1(q)^{-1}} dx \\ \int_{\mathbb{R}^{2d}} \langle p, r_2(x) \rangle dx &\mapsto -\tilde{\beta} \int_{\mathbb{R}^{2d}} \hat{\varrho}(x)|p|_{K_2(q)^{-1}}^2 dx \\ \int_{\mathbb{R}^{2d}} \langle \nabla V(q), r_2(x) \rangle_{K_2(q)} dx &\mapsto \frac{1}{\tilde{\beta}} \int_{\mathbb{R}^{2d}} \langle \nabla V(q), \nabla_q \hat{\varrho}(x) \rangle_{K_2(q)} dx.\end{aligned}\tag{5.44}$$

Substituting (5.41) and (5.42) into the expression (5.8) for $I_a(b)$, applying (5.44) and dropping terms independent of β and $V(q)$ gives the following functional of $(V(q), \beta)$:

$$\begin{aligned}I_a(\beta, V) &= -\frac{\beta^2}{4} \int_{\mathbb{R}^{2d}} \hat{\varrho}(x)|\nabla V(q)|_{K_1(q)^{-1}}^2 dx - \frac{\beta^2}{4} \int_{\mathbb{R}^{2d}} \hat{\varrho}(x)|p|_{K_2(q)^{-1}}^2 dx \\ &\quad - \frac{1}{4} \int_{\mathbb{R}^{2d}} \hat{\varrho}(x)|\nabla V(q)|_{K_2(q)}^2 dx - \frac{\beta}{2} \int_{\mathbb{R}^{2d}} \langle \nabla V(q), \nabla_q \hat{\varrho}(x) \rangle_{K_1(q)^{-1}} dx \\ &\quad + \frac{\beta\tilde{\beta}}{2} \int_{\mathbb{R}^{2d}} \hat{\varrho}(x)|p|_{K_2(q)^{-1}}^2 dx - \frac{1}{2\tilde{\beta}} \int_{\mathbb{R}^{2d}} \langle \nabla V(q), \nabla_q \hat{\varrho}(x) \rangle_{K_2(q)} dx\end{aligned}\tag{5.45}$$

This functional is quadratic in each of β and $\nabla V(q)$ separately. It may be written as

$$\begin{aligned}
 I_a(\beta, V) = & -\frac{1}{4} \int_{\mathbb{R}^{2d}} \hat{\rho}(x) |p|_{K_2(q)}^2 dx (\beta - \tilde{\beta})^2 \\
 & -\frac{1}{4} \int_{\mathbb{R}^{2d}} \left| \hat{\rho}^{\frac{1}{2}} \nabla V(q) + \frac{1}{\tilde{\beta}} \frac{\nabla_q \hat{\rho}(x)}{\hat{\rho}(x)^{\frac{1}{2}}} \right|_{K_2(q)}^2 dx \\
 & -\frac{\beta^2}{4} \int_{\mathbb{R}^{2d}} \left| \hat{\rho}(x)^{\frac{1}{2}} \nabla V(q) + \frac{1}{\tilde{\beta}} \frac{\nabla_q \hat{\rho}(x)}{\hat{\rho}(x)^{\frac{1}{2}}} \right|_{K_1(q)}^2 dx + \epsilon.
 \end{aligned} \tag{5.46}$$

where $\epsilon > 0$.

From this expression it is clear that the maximum is attained by choosing $\beta = \hat{\beta}$ and $V(q) = \hat{V}(q)$ where $\hat{\beta} = \tilde{\beta}$ and $\hat{V} = -\frac{1}{\tilde{\beta}} \log \hat{\rho}(x)$.

5.6 Estimating the Diffusion Coefficient

The discussion in sections 5.5 and 5.4 shows that it is possible to fit first and second-order Langevin equations to the empirical measure generated by a time-series. In doing so the model fit is completely specified, up to a time-constant k . In this section we show how standard practitioners' approaches to determining this time-constant, through fitting the auto-correlation function, are closely related to common statistical practise which focuses on finding the quadratic variation. We concentrate on the first order case, and assume that we are given a time series $\{x_n\}_{n=0}^{N-1}$. The second order case is similar.

5.6.1 Statisticians' Approach

The statisticians' approach is to fit the diffusion co-efficient using the quadratic variation. For the first order Langevin equation standard properties of diffusion processes show that $\frac{1}{k}$ can be estimated by the formula

$$2kI \approx \frac{1}{N\Delta t} \sum_{n=0}^{N-1} (x_{n+1} - x_n) \otimes (x_{n+1} - x_n). \tag{5.47}$$

5.6.2 Autocorrelation Function

The (unnormalised) autocorrelation function is defined by

$$C(\tau) = \mathbb{E}x(0)x(\tau)$$

on the assumption that $x(0)$ is distributed according to the invariant measure. It is known that $c'(0) = -\frac{1}{\beta}$. By ergodicity the auto-correlation function can be expressed as the time-average

$$C(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)x(t+\tau)dt.$$

For the first-order Langevin equation, practitioners will often estimate $\frac{1}{\beta}$ by adjusting it so that the autocorrelation function of the model best fits the data. For example one might fit the constant so that the slope $C'(0)$ fits the data. We show that this latter procedure is directly related to estimating the quadratic variation as above.

Rearranging the expression used to estimate $\frac{2}{\beta}$ in (5.47) we find that

$$\begin{aligned} [C(0) - C(\Delta t)] &= \frac{1}{N} \sum_{n=0}^{N-1} x_n^2 - \sum_{n=0}^{N-1} x_n x_{n+1} \\ &= \frac{1}{N} \sum_{n=0}^{N-1} (x_n^2 - x_n x_{n+1}) + \frac{x_N^2 - x_0^2}{2N} \\ &= \frac{1}{2N} \sum_{n=0}^{N-1} (x_{n+1}^2 + x_n^2 - 2x_n x_{n+1}) \\ &= \frac{1}{2N} \sum_{n=0}^{N-1} (x_{n+1} - x_n) \otimes (x_{n+1} - x_n). \end{aligned}$$

The first line is a natural approximation for the derivative of the auto-correlation function, expressed in terms of time-averages. Subsequent lines show that this approximation can be re-written in terms of the quadratic variation. Thus fitting the slope of the empirical auto-correlation, as practitioners do, is closely related to the standard statistical procedure of estimating the quadratic variation. Fitting various transforms of the auto-correlation, however, is more involved and exploits knowledge of the drift terms.

5.7 Relationship to existing literature

In the statistical literature, the estimation of diffusion parameters is usually viewed as straight-forward: In the case of continuous time trajectories the estimate is given using quadratic variation and a limiting process like the one in (3.3), this is noted e.g. in the foreword of [62]. For discrete time observations, an assumption of high frequency

observations is usually made and estimators tend to be based on quadratic variation at finite inter sample times like (3.13). Theorems on the asymptotic behaviour of these estimators are available, e.g. in [50].

In the chemistry and physics literature, estimating the diffusion coefficient has been viewed in conjunction with assessing a range of timescales where the diffusion process provides a good approximation of the true dynamics. Hummer notes in [34] that position dependent diffusivity should be employed which corresponds to multiplicative noise. He offers a fully Bayesian algorithm based on binning and then considering transition rates between bins. More traditional methods from this field reduce consideration to a small region around an equilibrium point in phase space, using a harmonic approximation for the potential. One can then fit analytic expressions for the Fourier spectrum of the velocity autocorrelation in the harmonic oscillator case, see [32] and [59], yielding the friction coefficient γ (and σ via fluctuation-dissipation). Alternatively (depending on whether the over- or under damped regime is considered) one can consider a fit to the Laplace transformation of the velocity autocorrelation as described in [36]. Even more traditionally, one can look at the mean square displacement of the particles and infer diffusivity. Again, if the drift parameters are known (or can be approximated well in the region of interest) one can identify the diffusivity from spatial autocorrelations.

Estimating the drift coefficients is generally considered easy as they can be inferred from the potential of mean force, as pointed out by Hummer, [34], although computing the potential of mean force in areas that are less well sampled poses a challenge that has led to a plethora of algorithms. The statistical literature typically considers drift parameter estimation to be the harder problem of the two. Methods akin to the practitioners' approach via counting population densities include in particular the minimum distance estimator. Kutoyants (see [62]) gives two realisations of the minimum distance estimator. Since the cumulative distribution function of the processes involved here is not explicitly available here, whereas the pdf is, we settle for Kutoyants' second minimum distance estimator which seeks to minimise the following functional:

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \|\hat{\rho}(\cdot) - \rho(\theta, \cdot)\|_{L^2} \quad (5.48)$$

Here, as above, $\hat{\rho}(\cdot)$ refers to the empirical density whereas $\rho(\theta, \cdot)$ denotes the invariant

probability density induced by our SDE for the choice of drift parameters θ .

The choice of L^2 -norm in this case is somewhat arbitrary, although it is of course useful, from an implementation point of view. The relationship of this minimum distance estimator to the practitioners' estimator studied in this paper is not as straightforward as it seems (see section 5.8) because the potential fitted by the practitioners in a least squares sense needs to be exponentiated, normalised and then fitted in an L^2 -sense.

The fact that θ will normally be finite-dimensional implies that some interpolation error will always be made, and this interpolation error is transformed by the nonlinearity of $\exp(\cdot)$ and L^2 -fitting in a nonlinear way.

5.8 Numerical Experiments

5.8.1 Introduction

As an example on which to perform experiments we choose the simple one-dimensional diffusion with a gradient vector field given by

$$dx = \left(-x^3 + \frac{3}{2}x\right) dt + 2.5dW, \quad x(0) = 0 \quad (5.49)$$

Using this special case of (5.13) we will study how the practitioners' method, the second minimum distance estimator as well as the maximum likelihood estimator perform when used to estimate drift parameters. The functional form to be fitted to this SDE is given as:

$$dx = \sum_{i=1}^c (\theta_i x^{i-1}) dt + \sigma dW, \quad x(0) = 0 \quad (5.50)$$

Here, both the θ_i as well as the diffusion coefficient σ are to be estimated. In order to distinguish contributions due to the $\mathcal{O}(1/T)$ term in (5.15) related to the initial condition and other contributing factors, a new drift estimator is introduced based on maximising the functional given in (5.15) without the first term:

$$\tilde{\theta} = \operatorname{argmin}_{\theta} \frac{1}{2T} \int_0^T (|\nabla V(x; \theta)|^2 - \frac{2}{\beta} \Delta V(x; \theta)) dt. \quad (5.51)$$

This estimator can be expressed explicitly in the case of linear parameter dependence, (5.50), that is at hand.

This section is organised as a collection of numerical experiments as follows:

1. A non-parametric view of the correlation between estimated potentials expected from the link between MLE and Practitioners' methods is presented in subsection 5.8.2, see figure 5.1. This serves as preliminary numerical confirmation of the practical relevance of the claimed link.
2. Parametric estimation based on the MLE, the practitioners' approach and the second MDE will be introduced in subsection 5.8.3 including numerical illustration of asymptotic consistency in figures 5.2 to 5.7.
3. The correlation structure of four parametric drift estimators (MLE, 2nd MDE, Practitioners', $\tilde{\Theta}$ -method) will be investigated in subsection 5.8.4 with figures 5.9 to 5.11 summarising the main results.
4. A brief note on comparing estimated parameters via their induced autocorrelations is made in subsection 5.8.5 giving results in figures 5.12 and 5.13. It shows that in the first order case induced autocorrelations may not be a very sensitive benchmark, whereas in the second order case it is more telling as shown in figure 5.14.
5. Finally, we will investigate in subsection 5.8.6 how these estimators perform in the case of a misspecified model where the sample paths are generated using multiplicative noise. As figure 5.15 shows, the invariant probability densities are badly reproduced by the MLE and fairly well reproduced by the second MDE while both estimators fail to produce the correct induced autocorrelations, see figure 5.16.

5.8.2 Empirical and MLE-induced Probability Densities

In order to broach the relation of maximum likelihood estimates for drift parameters and the empirical density produced by the process, we perform a few preliminary essentially non-parametric experiments and follow these up with a more careful study of correlation of estimated drift parameters using the different estimators.

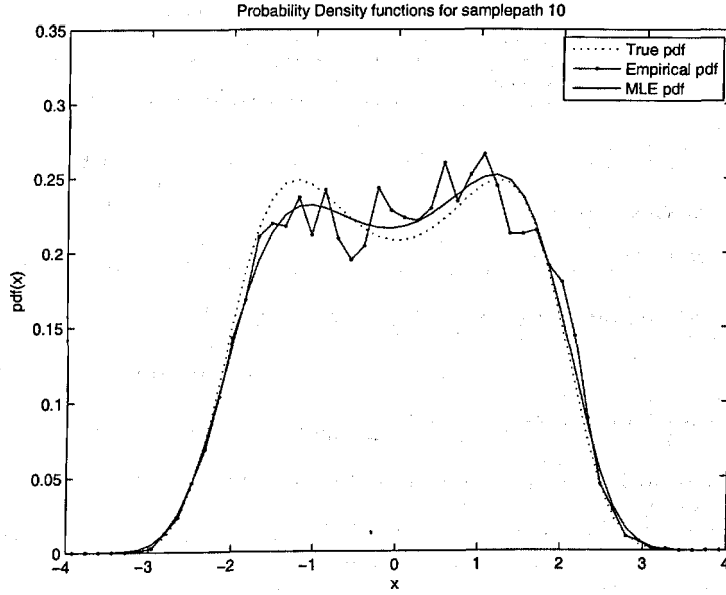


Figure 5.1: probability density functions from one particular samplepath

An ensemble of $N = 100$ sample paths for the SDE (5.49) is created for final time $T = 100$ and sampled at spacing $\Delta t = 0.01$. The paths are created using a subsampled ($k = 30$) Euler-Maruyama method. On each of these paths, a maximum likelihood estimator based on an Euler approximation is used for the parameters θ_i . Also, a histogram is computed for each of these paths using $B = 50$ bins spaced equidistantly on the interval $[-4, 4]$.

The true probability density function (pdf) as well as the pdf arising from the MLE-estimated parameters θ and the empirical pdf from the histogram are plotted in figure (5.1) in a typical case.

As it is difficult to derive from these graphs whether such an agreement is indeed typical or merely coincidental, a measure of correlation is computed as follows. At the centres of the bins, denoted by $\{c_i\}_{i=1,\dots,50}$, the deviations of the MLE-derived pdf for path i , $g_{MLE}^{(i)}$ from the true pdf, g , as well as the deviation of the empirical pdf, $g_{EMP}^{(i)}$ are computed to form the following correlation coefficient, summing over $N = 100$

realisations:

$$c = \sum_{i=1}^N \frac{\sum_{b=1}^B (g_{EMP}^i(c_b) - g(c_b)) \cdot (g_{MLE}^{(i)}(c_b) - g(c_b))}{\sqrt{\sum_{b=1}^B (g_{EMP}^i(c_b) - g(c_b))^2} \cdot \sqrt{\sum_{b=1}^B (g_{MLE}^{(i)}(c_b) - g(c_b))^2}}$$

The correlation coefficients obtained in two experiments with different random seeds were $c = 0.83$ and $c = 0.81$ respectively, so that some degree of correlation is present. However, the presence of terms relating to the initial and final conditions in (5.18) shows that no more should be expected. Also, errors due to finite Δt and finite number of bins B will play a role.

The (root mean square) average deviation in (5.18) will be of size $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ whereas the deviation due to the initial conditions is only $\frac{1}{T}$, so that increasing the final time T would be expected to improve this correlation, provided that effects due to finite Δt are negligible.

5.8.3 Parametric Estimation

Implementation of the Practitioners' method

We wish to adapt and implement the estimator given by (5.26) for the standard 1d example (5.49). Since the estimator is inherently non-parametric whereas the model to be fitted, (5.50), is parametric, some adaptation is needed.

This comes in the form of first computing the histogram (based on a number of bins B (usually 50)) on the interval $[4, 4]$. Using this histogram data and the quadratic variation estimator for σ in (5.49), a least-squares fit is performed so that (5.26) is satisfied approximately in a least error squares sense given the functional form, (5.49) to be fitted:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \sum_{b=1}^B \left| \frac{\sigma^2}{2} \log(\varrho(c_b)) - \sum_{i=1}^c \theta_i V_i(c_b) \right|^2 \quad (5.52)$$

It should be pointed out that any finite choice of the number of bins B is likely to incur an error in estimated parameters as the choice of the bin centre c_b for evaluation of V_i is arbitrary. To be accurate, this would have to be replaced by an appropriately (logarithmically) weighted integral of this function over the bin interval.

A further problem with this method arises when some parts of the interval $[-4, 4]$ are poorly sampled. This poor sampling results in a jagged logarithmic histogram and large deviations from any accessible invariant density in the parameter space spanned by the $\{\theta_i\}$ seem to negatively affect the estimator. To mitigate this problem, a cutoff is introduced whereby only those bins which contain at least $\frac{1}{10}$ of the samples to be expected under uniform distribution are taken into account in the least squares fitting. Practitioners might well introduce a weighting for the errors to mitigate these effects and a mathematically more sophisticated approach might use the Kullback-Leibler divergence (relative entropy) to obtain the estimate. This, however, is computationally slightly more cumbersome and we feel that for simple illustration purposes, this ad-hoc criterion performs well without unduly affecting the core performance of the estimator.

Implementation of the second MDE

In order to compare MDE and MLE estimates of the drift coefficients, a parametric minimum distance estimator is required. Kutoyants (see [62]) gives two variants of the minimum distance estimator. Since the cumulative distribution function of the processes involved here is not explicitly available, whereas the pdf is, we settle for Kutoyants' second minimum distance estimator which seeks to minimise the following functional:

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \|g_T(\cdot) - f(\theta, \cdot)\|_{L^2} \quad (5.53)$$

Here, $g_T(\cdot)$ refers to the empirical density whereas $f(\theta, \cdot)$ denotes the invariant probability density induced by our SDE for the choice of drift parameters θ .

In general, the minimum need not be attained and there are no guarantees that it is unique, either. We therefore search for a local minimum and apply a steepest descent algorithm in a 50-bin discretisation of the pdfs. The termination criterion is for the gradient of the functional wrt. θ to be below a certain threshold. Convergence is observed in a large majority of cases, occasionally, if the histogram is very jagged, the algorithm grinds to a halt.

The input value for the diffusion coefficient σ to this algorithm is computed from the quadratic variation of the input path using the estimator (3.13). This is the only

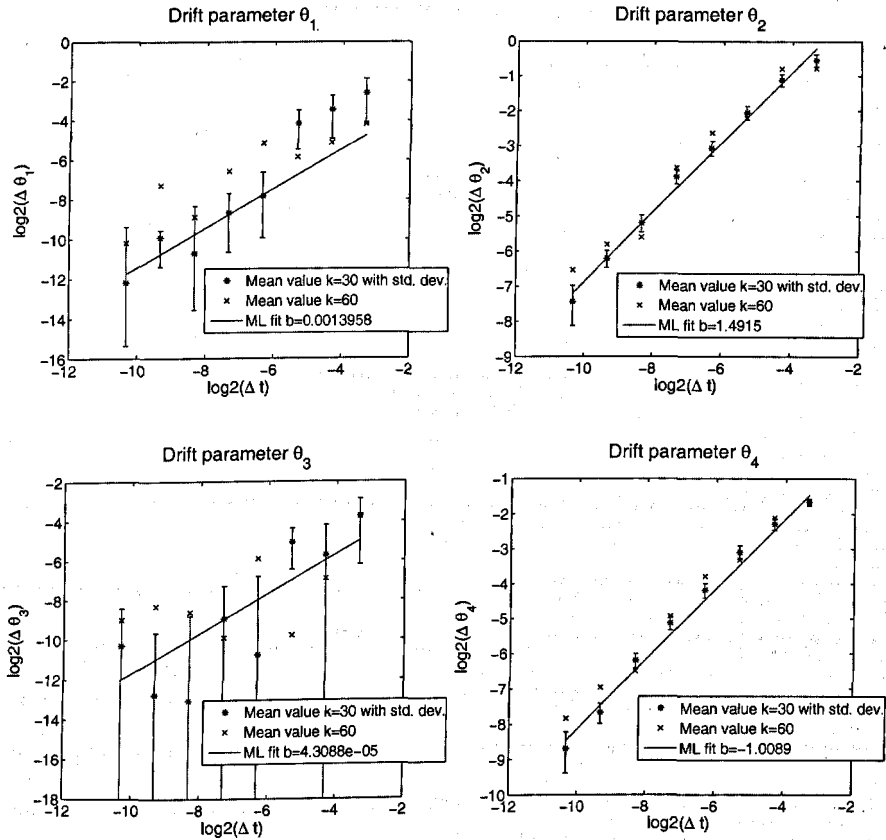


Figure 5.2: Convergence as $\Delta t \rightarrow 0$ of drift parameters for MLE

part where effects resulting from finite inter sample times Δt enter the estimator.

Asymptotic consistency

To demonstrate asymptotic consistency of these estimators, we compute MDE, Practitioners' and MLE estimates of drift parameters using the SDE 5.49. Performing this with a final time of $T = 100$ and timesteps from $\Delta t \in \{0.1, \frac{0.1}{2}, \dots, \frac{0.1}{128}\}$ we obtain the plots given in figures 5.2 and 5.3 for the MLE estimates using the same linear error model as in (4.48).

The estimate for the asymptotic drift coefficients is not expected to be consistent (statistical significance is open for now) since there is error related to finite final time in

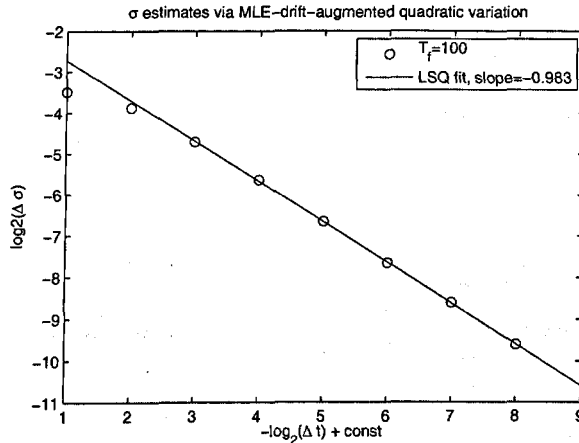


Figure 5.3: Convergence as $\Delta t \rightarrow 0$ of diffusion parameter for MLE

conjunction with starting all sample paths from the starting point $x(0) = 0$. Error due to finite Δt is imported into the MDE estimates via the σ estimate which is why figures 5.4 and 5.5 show Δt related error. Those drift parameters which are zero (i.e. θ_1 and θ_3) show very small estimation error indeed. Since this error is not necessarily caused by the incorrect σ (and hence not multiplicative, $\mathbb{E}\hat{\theta} = \theta(1 + \mathcal{O}(\Delta t))$) but more likely to be due to the finite number of bins, it is understandable that the linear error model cannot convincingly account for Δt related error. Concerning this graph it should also be pointed out that the confidence intervals sometimes extend all the way to $-\infty$ which is simply due to the extrapolated (for $\Delta t \rightarrow 0$) value for the drift parameter being inside the confidence interval, so that the set distance between the confidence interval and the extrapolated value is zero, corresponding to the logarithm $-\infty$.

It should be noted that, similarly to the practitioners' estimator, choosing a constant number $B = 50$ of bins for the MDE is questionable. However, given perfect histogram data (attained by artificially setting the histogram entries to the values of the pdf generated by the correct values for drift and diffusion parameters) the MDE as implemented has been observed to converge to the true drift parameter values (to within ± 0.004) from 100 randomly sampled starting conditions, see figure 5.6.

Finally, the drift parameters graph obtained for the practitioners' estimator is

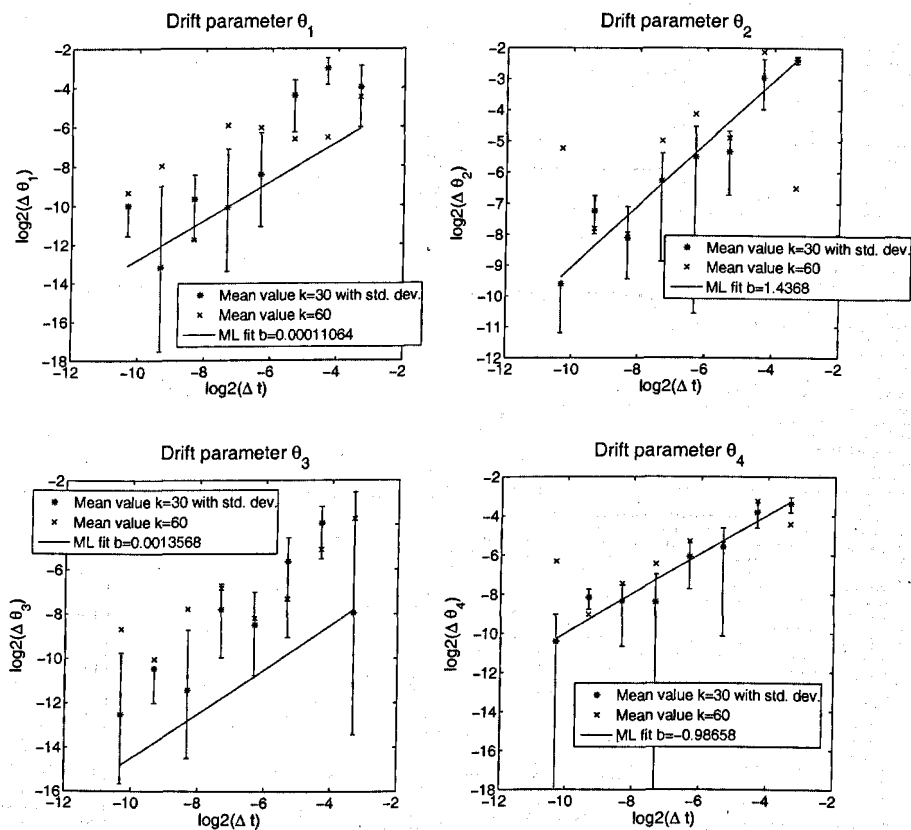


Figure 5.4: Convergence as $\Delta t \rightarrow 0$ of drift parameters for MDE

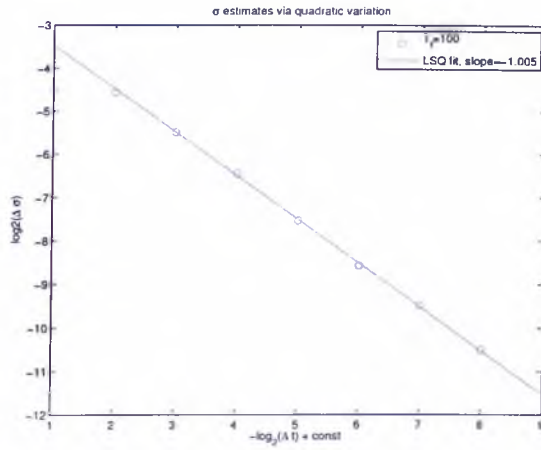


Figure 5.5: Convergence as $\Delta t \rightarrow 0$ of diffusion parameter for MDE

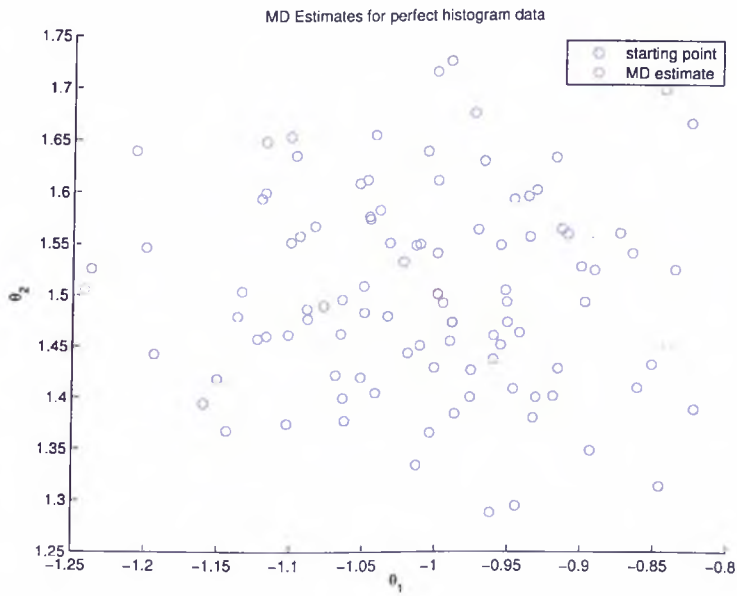


Figure 5.6: Confluence of MD estimates for perfect histogram data

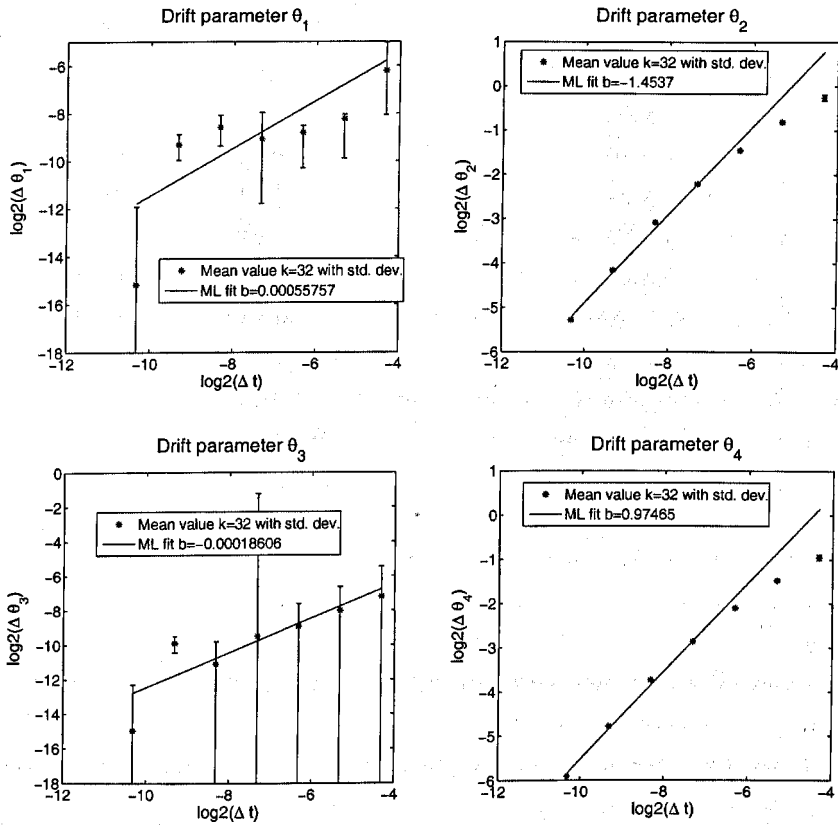


Figure 5.7: Convergence θ_i as $\Delta t \rightarrow 0$ of drift parameters for Practitioners' estimator

displayed in figure 5.7, a separate display for the consistency of σ is redundant as $\hat{\sigma}$ is arrived at using the estimator (3.13) based on quadratic variation alone which was previously shown to be asymptotically consistent in this implementation. Similar comments as to small errors in θ_1 and θ_3 as well as confidence intervals extending to $-\infty$ in the logarithmic display apply.

5.8.4 Correlation of estimated drift parameters

The experiment in 5.49 is repeated with diffusivity $\sigma = 1.5$, final times

$T_f \in \{10, 20, 40, \dots, 10240\}$ and timestep sizes $\Delta t \in \{2 \cdot 10^{-2}, 2 \cdot 10^{-3}, 2 \cdot 10^{-4}, 2 \cdot 10^{-5}\}$.

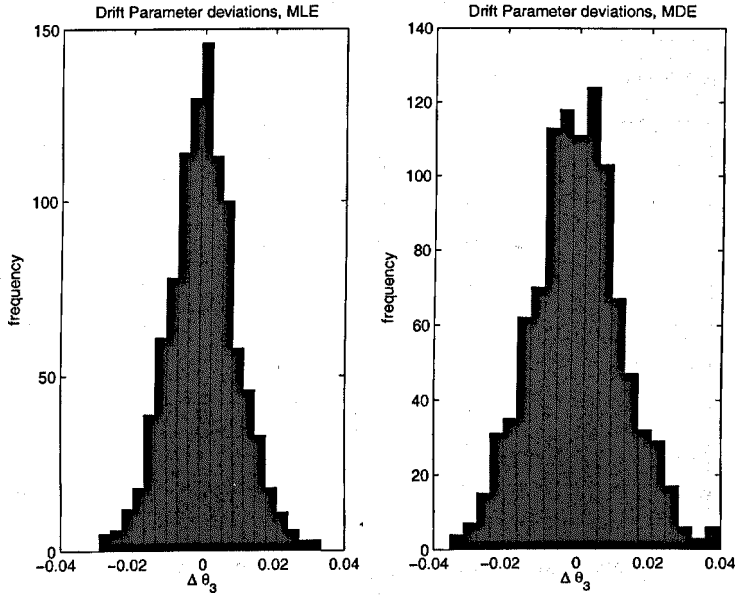


Figure 5.8: Deviations of drift Parameter θ_3 from mean, $T_f = 20480$

Using $N_{av} = 500$ sample paths for each configuration, the MLE and MDE estimates are computed. It is found that their variances do indeed decay like $\mathcal{O}(\frac{1}{T})$. The deviations of the estimators from their respective means (over fixed T_f) is computed. These deviations display an approximately Gaussian distribution, $\mathcal{N}(0, \frac{\text{const}}{T})$, as shown in figure 5.8 for $T_f = 20480$ and $\Delta t = 0.02$.

Plotting the averaged correlations as a function of final time T_f yields the plot in figure 5.9.

It seems that the maximal obtainable correlation coefficient is around 0.7. For small final times T_f , the influence of errors related to finite resolution Δt is apparent and an increase in observed autocorrelation with resolution is clear. For larger final times, however, the increase is not maintained.

It may be hypothesised that the $\tilde{\Theta}$ estimator should be more correlated with the MDE since it is based on performing the same Stratonovich/integration process as above. In fact, the de-correlation of the $\tilde{\Theta}$ and MLE should indicate the influence of the initial-condition related term in (5.15) on the parameter estimates. We compute

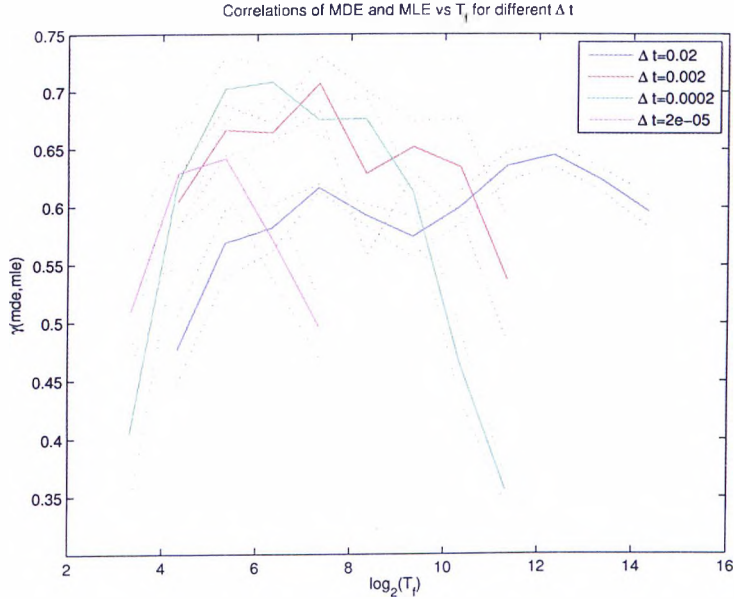


Figure 5.9: Correlations of drift parameter deviations

the correlation of the \bar{A} estimate and the MLE for the same drift parameter as above, again using $T_f \in \{10, 20, 40, \dots, 10240\}$ and $\Delta t = 0.0002$ in this case, which results in figure 5.10.

The remarkably high degree of correlation indicates that the first term which is of order $\mathcal{O}(\frac{1}{T})$ is of little influence.

The main reason for the correlation not approaching 1 in figure 5.9 must thus be sought elsewhere. Since the discretisation influence exerted by finite Δt as well as scaling with final time T_f have been investigated and do not seem to account for all of the deficiency, other potential culprits may include finite numbers of bins in the histogram. However, this is unlikely given the observed asymptotic consistency of the second MDE.

While the variational characterisation of the optimal fitted potential as a critical point of the functional connected to $\bar{\theta}$ is correct, the second MDE does *not* provide a potential which is a critical point of that functional. The projection process involves ad-hoc choices such as an L^2 norm as well as a choice of basis for the space of potentials.

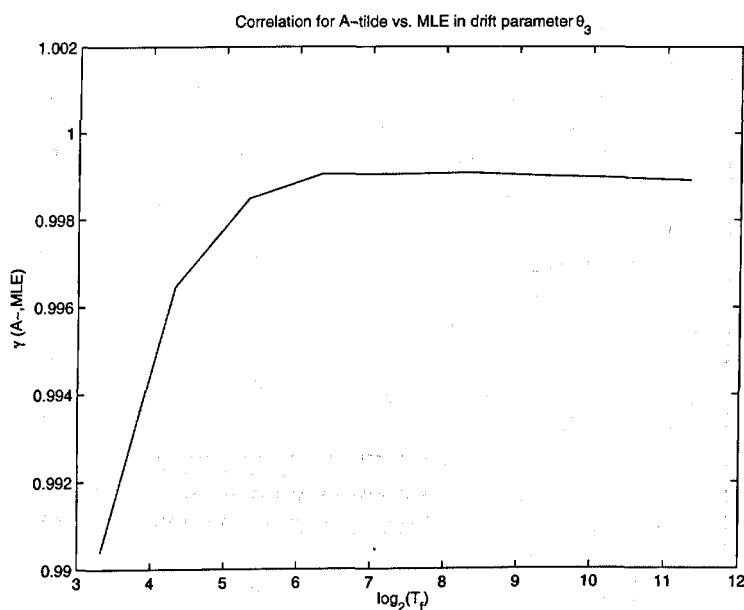


Figure 5.10: Correlations of drift parameter deviations for \tilde{A} vs. MLE

It would be interesting to see whether, as more and more basis vectors are included, the maximal observed correlation is increased. In view of the ill-conditioning and it being questionable whether there would be sufficient decay of the ('Fourier') coefficients of the potential a numerical investigation of this question seems hopeless.

To further elucidate the question whether the low observed correlations between estimated drift parameters should be attributed to uncorrelated interpolation error due to finite polynomial order or exponentiation and the choice of the L^2 -norm, correlations with drift parameters estimated using the Practitioners' estimator, (5.52) are examined. Using a constant timestep Δt and a range of final times as above it can be seen from figure 5.11 that the Practitioners' estimator and the MLE are more strongly correlated than any other pair of estimators. While some caution has to be exercised as this experiment was conducted only for one size of timestep and in view of the rather large standard deviations, it would seem that the source of de-correlation is in fact related to exponentiation and L^2 -norm rather than interpolation error.

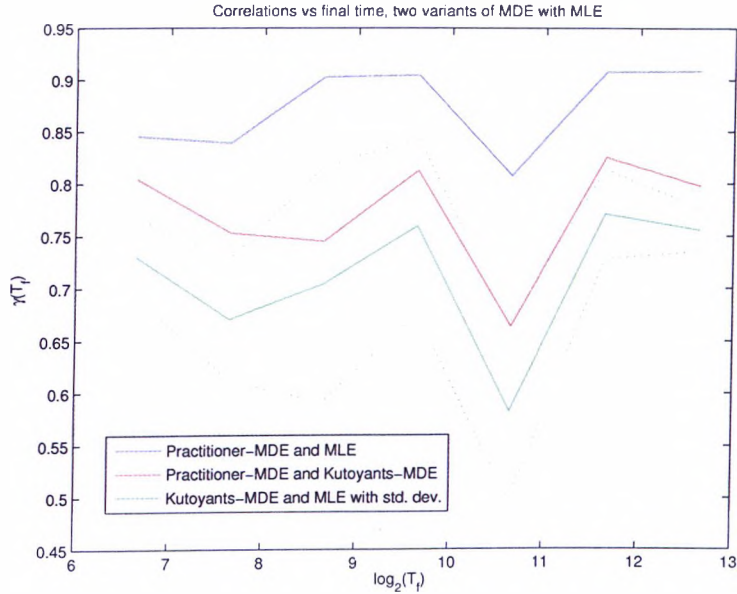


Figure 5.11: Comparing correlations of estimated drift parameters for MDE, Practitioners' and MLE estimated drift parameters

5.8.5 Comparing autocorrelations

In order to compare MDE and MLE some 'independent' yet meaningful statistical test would be helpful. Kutoyants ([62]) points out that the MLE is best (asymptotically efficient) at reproducing the likelihood integral (of the kind $\int_0^T \left| \frac{dW}{dt} \right|^2 dt$ with appropriate interpretation relative to Wiener measure), whereas the MDE is best at reproducing the histogram for slightly contaminated models. This is hardly surprising.

On the other hand, it seems clear that in fitting 'slightly' misspecified models (e.g. multiplicative noise for path sampling vs. additive noise parameter fitting, see subsection 5.8.6) there will always be some statistical test which the fitted model will not pass (e.g. binning of local quadratic variation showing statistically significant differences of the fitted model from the supplied path). If fitting is to be used as a means of establishing parameters that 'work best' if used in a simplified model of a real system, then some *practically meaningful* benchmark of how the fitted model is doing is more helpful than a contrived statistical test aiming only at highlighting its specific deficiencies.

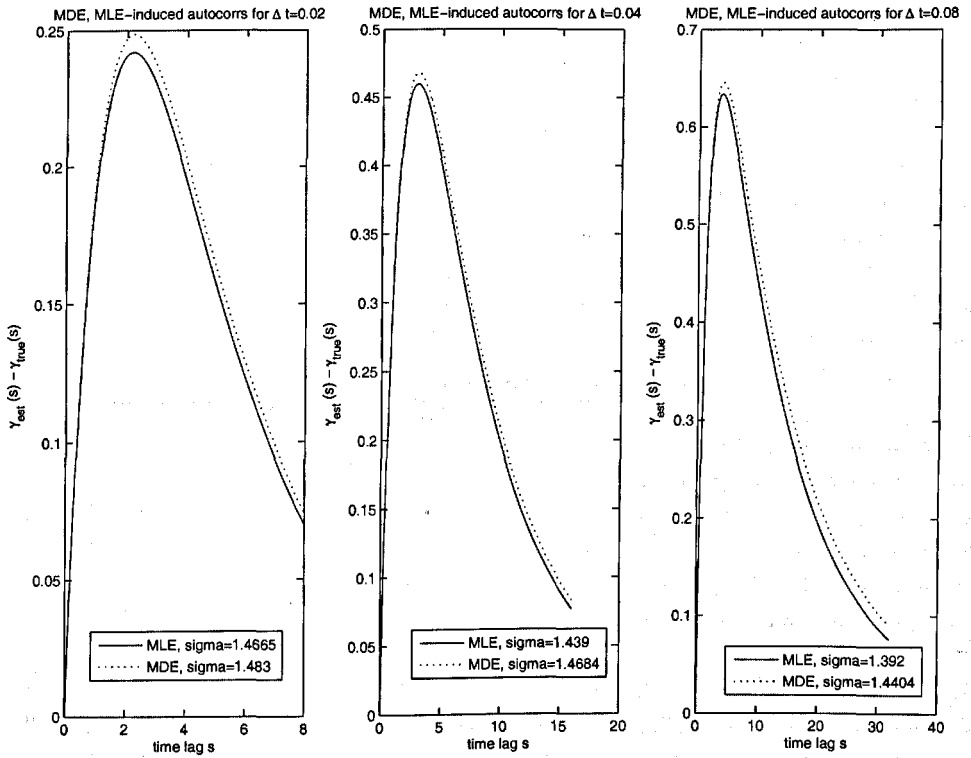


Figure 5.12: Comparing autocorrelations for MDE and MLE estimated parameters

Such a practically meaningful benchmark may be given by the induced autocorrelation. This comparison is performed for the same experimental setup as above, using $\sigma = 1.5$, otherwise as given in (5.49). Figure 5.12 gives plots of the difference of the autocorrelations obtained for the MDE-estimated and MLE-estimated drift and diffusion coefficients. It appears that the main difference is induced by incorrectly estimated diffusion coefficients.

Performing the same experiment giving the MDE the true value of σ yields autocorrelations for the MDE that were found to be on the verge of being statistically significantly different from the true autocorrelations only at considerable CPU cost. Note that the data presented in figure 5.13 does not take into account the deterministic numerical error in computing the autocorrelations.

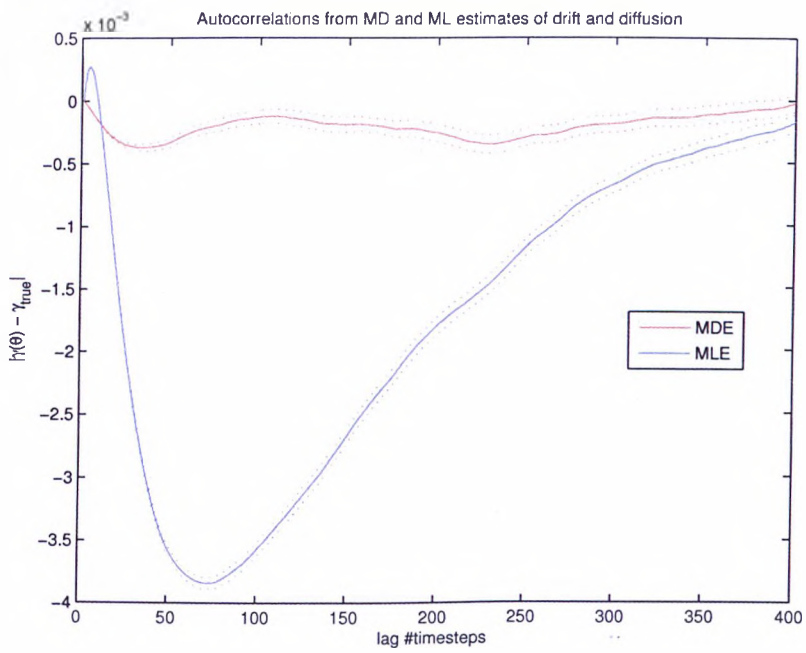


Figure 5.13: Comparing autocorrelations for MDE and MLE estimated parameters, true σ for MDE

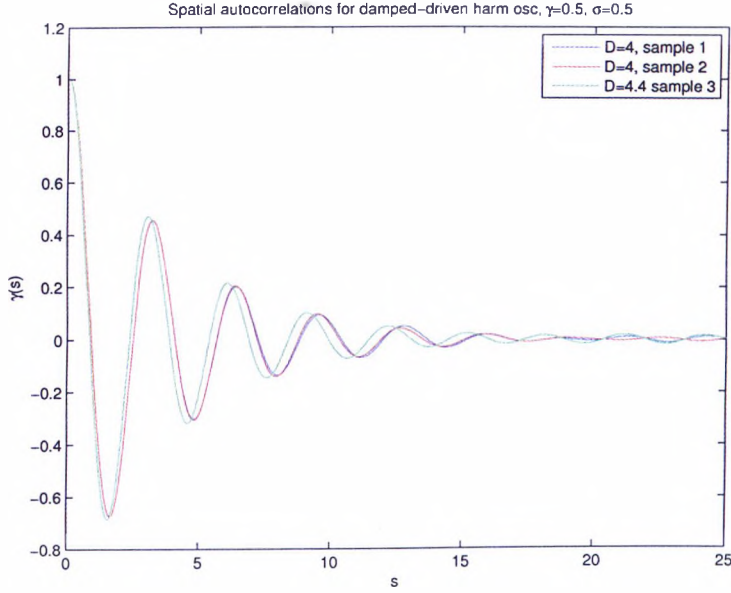


Figure 5.14: Autocorrelations for damped-driven harmonic oscillator

As comparing autocorrelations may be more interesting in the second order case, we consider a damped-driven harmonic oscillator as described by the following SDE:

$$\begin{aligned} dq &= p dt \\ dq &= (-Dq - \gamma p) dt + \sigma dB \end{aligned} \quad (5.54)$$

In order to establish identifiability, numerical autocorrelations are computed employing one sample path each generated using the following parameter values:

$$T = 5 \cdot 10^4 \quad \Delta t = 0.05 \quad \gamma = 0.5 \quad \sigma = 0.5 \quad D \in \{4, 4.4\}$$

The autocorrelations obtained for three different random seeds are displayed in figure 5.14. It is clear that deviations of the drift parameters can be discerned using the autocorrelation so that a meaningful comparison of MDE and MLE might be possible. The form of the deviation is easily understood: A higher value for D means going further into the under damped regime, yielding higher correlations overall as well as a faster eigenfrequency so that a time-lag rescaling occurs.

It is apparent that spatial autocorrelation provides a means of distinguishing

different sets of drift parameters in the second order case and could thus be used to benchmark fitted SDEs. In some way, this is the reverse of the approach of [32], [59] and many others, where the (velocity) autocorrelation is used to estimate drift parameters. It is clear that much more can be done concerning the second order case which has been central to many practical applications.

5.8.6 Misspecified model – multiplicative noise

Motivated by observations made fitting scalar (1d) second order Langevin-type SDEs with trigonometric polynomial potential and additive noise to Langevinised MD simulation data in chapter 4 we investigate robustness of MLE and MDE against misspecified models.

To create an even simpler example exhibiting the main problem in the aforementioned application, we consider the following SDE:

$$dx = \left(-x^3 + \frac{3}{2}x \right) dt + \frac{3}{\sqrt{4 + (1.22 - x)^2}} dB, \quad x(0) = 0 \quad (5.55)$$

This corresponds to a higher 'temperature' at the right equilibrium than at the left equilibrium point, so while it does not change the sampled drift parameters when an MLE is used, it greatly changes the histogram (particles spending less time at the 'hot' equilibrium).

Using the parameters $T_f = 80$, $\Delta t = 0.002$ and $k = 32$, the average drift parameter (over $N = 400$ realisations) and diffusion parameters θ and σ induce pdfs displayed in figure 5.15.

It is clear that the MDE estimated drift parameters are far better at reproducing the true pdf, even though it is apparent that the variation in the induced pdf is a bit too large. This can be traced back to a slightly underestimated σ , which would be expected to improve with increased temporal resolution.

It should be stressed that when fitting misspecified models, one can always find a statistic which the fitted model does not reproduce correctly, just that it happens to be the histogram (and hence the fitted potential) is inconvenient in a physically relevant case. Given its construction, it is clear that the MDE would be expected to

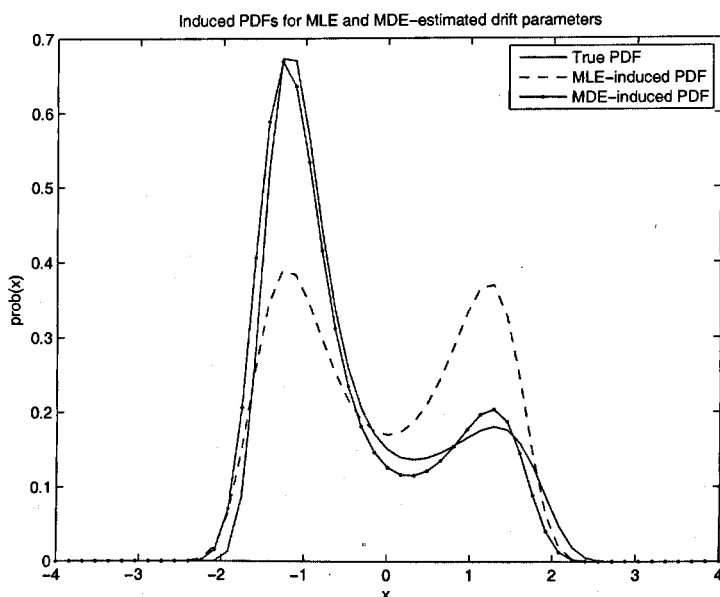


Figure 5.15: Comparing induced PDFs for MLE and MDE

yield misspecified fitted models which reproduce the invariant pdf better, however it would be expected to be bad at minimising the true likelihood. (analogous to formula (2.81) of [62] being optimised by the MLE rather than MDE). Having derived the MLE from the statistical model (3.11) in chapter 3 it is equally clear that this estimator will tend to reproduce the correct drift parameters. In fact, given a few extra technical hypotheses, theorem 1 still holds in this case so that asymptotically in the limit $\Delta t \rightarrow 0$ and $T_f \rightarrow \infty$ the true drift parameters will be recovered. In the presence of additive noise, however, these will lead to the wrong invariant distribution.

It is worth noting that this supports the contention that the failure of the estimator presented in chapter 4 to reproduce the invariant density for the Langevinised butane molecule (see figure 4.13) is due to the presence of multiplicative noise.

The autocorrelations for this misspecified model are not well-reproduced by either of the estimators, as can be seen from figure 5.16, where $N = 100$ realisations have been used.

The fact that MLE-estimated potentials yield faster de-correlation than MDE-

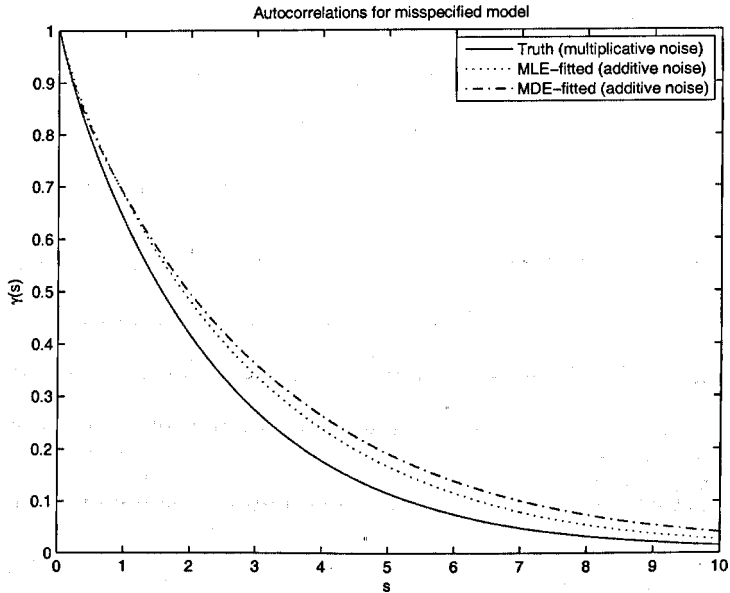


Figure 5.16: Comparing induced autocorrelations for MLE and MDE

estimated potentials is attributable to the deep well in the MDE-estimated potential in which paths can get 'stuck', resulting in strong correlation.

Also, the fact that the initial slopes of all three autocorrelations agree is reassuring, since this measure small-scale, diffusion-dominated de-correlation. Since the fitted (constant) diffusivity can be viewed as an ergodic average of diffusion over all paths, and even in the multiplicative noise case, variation of diffusivity is small over small time spans (smooth $\sigma(\cdot)$), this is expected.

5.9 Conclusions and Future Work

Significant analytical links between the maximum likelihood estimator used widely in the statistical literature and the Practitioners' estimator based on counting population densities have been found and studied on selected numerical examples. In the special case of gradient diffusions these estimators are even more closely linked as their deviations from the mean value satisfy the same statistics to leading order. While the minimum

distance estimator initially seems to be very close to the practitioners' approach, this turned out not to be accurate. Other links have been found between the statisticians' approach of estimating diffusivity via quadratic variation and the practitioners' reliance on fitted autocorrelations, although these are less close.

This chapter leaves open many avenues of further enquiry:

- Fitting diffusion coefficients given the drift parameters using only $\mathcal{O}(1)$ spaced data for a restricted class of models to be fitted might prove interesting from a statistical perspective. This could be built into a full sampling algorithm sampling alternately from drift and diffusion parameters.
- More consideration should be given to multiplicative noise models as applications are otherwise restricted to near-equilibrium configurations.
- A characterisation of the class of stochastic processes for which the link between MLE and the practitioners' method can be established would be desirable. Generalising from gradient diffusions to reversible processes is a first step. It is unclear, however, whether there is a more general class that would also include the second order Langevin process.
- It would be interesting to perform estimation for processes involving coloured noise such as

$$\ddot{q} + \nabla V(q) = B\dot{r}$$

where r is a suitable m -dimensional Ornstein-Uhlenbeck process involving \dot{q} to satisfy energy balance. Set up correctly, the process (q, \dot{q}, r) can have a product measure similar to the second order case.

Bibliography

- [1] Y. Ait-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation. *Econometrica*, 70(1):223–262, 2002.
- [2] B. M. Bibby and M. Sørensen. On estimation for discretely observed diffusions: A review. *Theory of Stochastic Processes*, 2(18):49–56, 1996.
- [3] B. Oksendal. *Stochastic differential Equations. An Introduction with Applications*. Springer, Berlin, 2000.
- [4] A. Le Breton and M. Musiela. Some parameter estimation problems for hypoelliptic homogeneous gaussian diffusions. *Seq. Meth. in Stat.*, 22:337–356, 1985.
- [5] M.E. Parker C. Xiao, D.M. Heyes. Cavitation in liquids by classical nucleation theory and molecular dynamics simulations, in *a.r.imre et al. (eds.)*, liquids under negative pressure. pages 231–242, 2002.
- [6] D. E. Catlin. *Estimation, Control and the Discrete Kalman Filter*. Springer-Verlag, 1989.
- [7] D.T. Crommelin and E. Vanden-Eijnden. Fitting timeseries by continuous-time markov chains: A quadratic programming approach. *J. Comp. Phys.*, page accepted for publication, 2006.
- [8] D.T. Crommelin and E. Vanden-Eijnden. Reconstruction of diffusions using spectral data from timeseries. *Comm. Math. Sci.*, page submitted, 2006.
- [9] C.W. Gardiner. *Handbook of Stochastic Methods*. Springer, Berlin, 1985.

- [10] A. M. Stuart D. Givon, R. Kupferman. Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, 17:55–127, 2004.
- [11] D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4):263–284, 1986.
- [12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data with the EM algorithm. *J. R. Stat. Soc., Ser. B*, 39(1):1–38, 1977.
- [13] R. Durrett. *Stochastic Calculus - A practical Introduction*. CRC Press, London, 1996.
- [14] G. Wanner E. Hairer, C. Lubich. *Geometric Numerical Integration*. Springer, 2002.
- [15] Y. Pokern E. Vanden-Eijnden, A. Stuart. Nonparametric estimation for diffusion processes. *in preparation*, 2006.
- [16] A. Beskos et al. Exact and computationally efficient estimation for discretely observed diffusion processes. *J. R. Statist. Soc. B*, 68(2):1–29, 2006.
- [17] B.R.Brooks et al. Charmm: A program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.
- [18] M. Galassi et al. *Gnu Scientific Library Reference Manual (2nd Ed.)*, release 1.4. <http://www.gnu.org/software/gsl/>.
- [19] M. Levitt et al. Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *J.Phys.Chem.B*, 101:5051–5061, 1997.
- [20] S.J.Weiner et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:65–784, 1984.
- [21] W.H.Press et al. *Numerical recipes in C : the art of scientific computing*. CUP, 1992.

- [22] A. W. Burgess F. A. Momany, R. F. McGuire and H. A. Scheraga. Energy parameters in polypeptides. geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Chem. Phys.*, 79:2361–2381, 1975.
- [23] A. Fischer. *Die Hybride Monte-Carlo-Methode in der Molekülphysik*. Diplomarbeit, Institut für Mathematik und Informatik, FU Berlin, 1997.
- [24] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989.
- [25] A. M. Stuart G. A. Pavliotis. *An introduction to Multiscale Methods, lecture notes*. 2006.
- [26] A. M. Stuart G. A. Pavliotis. Parameter estimation for multiscale diffusions. *submitted*, 2006.
- [27] M. Kac G. Ford. On the quantum langevin equation. *J. Stat. Phys.*, 46:803–810, 1987.
- [28] J. G. Gaines and T. J. Lyons. Variable step size control in the numerical solution of stochastic differential equations. *SIAM J. Appl. Math.*, 57(5):1455–1484, 1997.
- [29] V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. Henri Poincaré*, 29(1):119–151, 1993.
- [30] P. Giannopoulos and S. J. Godsill. Estimation of car processes observed in noise using bayesian inference. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, available from http://www-comserv.eng.cam.ac.uk/%7Esjg/pubs/pubs_noabst.html, 2001.
- [31] H. Grubmüller. *Molekulardynamik von Proteinen auf langen Zeitskalen*. Dissertation, Fakultät für Physik, TU München, 1994.
- [32] P.Tavan H.Grubmüller. Molecular dynamics of conformational substates for a simplified protein model. *J.Chem.Phys.*, 101:5047–5057, 1994.

- [33] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 1996.
- [34] G. Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(34), 2005.
- [35] A. Fischer I. Horenko, E. Dittmer and C. Schütte. Automated model reduction for complex systems exhibiting metastability. *Mult. Mod. Sim.*, to appear, 2005.
- [36] B.J.Berne J.E.Straub, M.Borkovec. Calculation of dynamic friction on intramolecular degrees of freedom. *J. Phys. Chem.*, 91:4995–4998, 1987.
- [37] J.S.Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, Berlin, 2001.
- [38] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, March 1960.
- [39] R. Khasminskii, N. Krylov, and N. Moshchuk. On the estimation of parameters for linear stochastic differential equations. *Probab. Theory Related Fields*, 113(3):443–472, 1999.
- [40] P. E. Kloeden, E. Platen, H. Schurz, and M. Sørensen. On effects of discretization on estimators of drift parameters for diffusion processes. *J. Appl. Prob.*, 33:1061–1076, 1996.
- [41] H. A. Kramers. *Physica*, 7(284), 1940.
- [42] D. Williams L. C. G. Rogers. *Diffusions, Markov Processes and Martingales, Vol.2*. CUP, 2000.
- [43] A. Lasota and M. C. Mackey. Springer, 1994.
- [44] D. M. Heyes M. E. Parker. Molecular dynamics simulations of stretched water: Local structure and spectral signatures. *J.Chem.Phys.*, 108:9039–9049.
- [45] X.-L. Meng and D. van Dyk. The EM algorithm—an old folk-song sung to a fast new tune. *J. R. Stat. Soc., Ser. B*, 59(3):511–567, 1997.

- [46] T.Nishimura M.Matsumoto. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8 No. 1:3–30, 1998.
- [47] D.J.Tildesley M.P.Allen. *Computer Simulation of Liquids*. OUP, 1987.
- [48] D. Nualart. *The Malliavin Calculus and Related Topics*. Springer-Verlag, 1991.
- [49] A. R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, 22:55–71, 1995.
- [50] B.L.S. Prakasa Rao. *Statistical Inference for Diffusion Type Processes*. Arnold Publishers, London, 1999.
- [51] H. Risken. *The Fokker Planck Equation*. Springer, 1984.
- [52] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.
- [53] G. O. Roberts and O. Stramer. On inference for nonlinear diffusion models using the hastings-metropolis algorithms. *Biometrika*, 88(3):603–621, 2001.
- [54] T. Schlick. *Molecular Modeling and Simulation, an Interdisciplinary Guide*. Springer, New York, 2002.
- [55] I. Shoji and T. Ozaki. Comparative study of estimation methods for continuous time stochastic processes. *J. Time Ser. Anal.*, 18(5):485–506, 1997.
- [56] G.O. Roberts S.K. Sahu. On convergence of the em algorithm and the gibbs sampler. *Statistics and Computing*, 9:55–64, 1999.
- [57] M. Mezei T. Schlick, S. Figueroa. A molecular dynamics simulation of a water droplet by the implicit-euler/langevin scheme. *J.Chem.Phys.*, 94:2118–2129, 1994.
- [58] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, 1997.
- [59] K. Schulten W. Nadler, A. T. Brünger and M. Karplus. Molecular and stochastic dynamics of proteins. *Proc. Natl. Acad. Sci.*, 84:7933–7937, 1987.

- [60] X.-G. Liang Y.-K. Guo, Z.-Y. Guo. Three-dimensional molecular dynamics simulation on heat propagation in liquid argon. *Chin.Phys.Lett.*, 18:71–73, 2001.
- [61] P.Wiberg Y. Pokern, A.M.Stuart. Parameter estimation for partially observed hypo-elliptic diffusions. *submitted to J. Roy. Stat. Soc.*, 2006.
- [62] Y.A.Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer, 2004.
- [63] R. Zwanzig. Nonlinear generalised langevin equations. *J. Stat. Phys.*, 9:215–220, 1973.