Maximum Likelihood Estimation for Stochastic Differential Equations Using Sequential Kriging-Based Optimization

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Grant W. Schneider, B.S., B.S., M.S.

Graduate Program in Statistics

The Ohio State University

2014

Dissertation Committee:

Peter Craigmile, Ph.D., Advisor
Radu Herbei, Ph.D., Advisor
Laura Kubatko, Ph.D.
Abstract

Stochastic differential equations (SDEs) are used as statistical models in many disciplines. However, intractable likelihood functions for SDEs make inference challenging, and we need to resort to simulation-based techniques to estimate and maximize the likelihood function. While sequential Monte Carlo methods have allowed for the accurate evaluation of likelihoods at fixed parameter values, there is still a question of how to find the maximum likelihood estimate. In this dissertation we propose an efficient Gaussian-process-based method for exploring the parameter space using estimates of the likelihood from a sequential Monte Carlo sampler. Our method accounts for the inherent Monte Carlo variability of the estimated likelihood, and does not require knowledge of gradients. The procedure adds potential parameter values by maximizing the so-called expected improvement, leveraging the fact that the likelihood function is assumed to be smooth. Our simulations demonstrate that our method has significant computational and efficiency gains over existing grid- and gradient-based techniques. Our method is applied to modeling stock prices over the past ten years and compared to the most commonly used model.
This dissertation is dedicated to Marceline Schneider, who passed away February 18, 2014. She was one of the strongest and most intelligent women I’ve ever had the privilege of knowing. The lessons I learned from her, particularly on the value of an education, were critical in the completion of this Ph.D. and to the person I’ve become.

Also, I dedicate this dissertation to my loving and supportive parents, Bill and Judy Schneider. My appreciation for their endless support and unconditional love could never be fully expressed in words.
Acknowledgments

I would like to thank all of the people who have helped and supported me throughout the past two decades or so of my education. Thank you to all of the wonderful teachers I had in the Frontier Local School District. In particular, I’d like to acknowledge and remember the three who have passed away since I’ve been at Ohio State: Mrs. Barbara Broffman, Mrs. Laura Hoff, and Ms. Marcia Ramsey. Thank you to Professor Richard Young and all the great professors I had as an undergraduate here at Ohio State University in the Honors Accounting program. Also thank you to Professor Ingrid Werner in the Finance Department. I cannot thank my advisors Professor Peter F. Craigmile and Professor Radu Herbei enough for their infinite patience and invaluable guidance over the course of my graduate work. I’ve learned so much from both of them, and I couldn’t have even begun to do this without them. Thank you to the entire Ohio State Statistics Department, in particular Professor Douglas Wolfe. For reasons I’m still not sure of, he trusted me to develop the R package to accompany his classic Nonparametric Statistical Methods book. A bit overwhelming at first, as I wasn’t sure I was qualified, it was an amazing learning experience and helped prepare me to begin on my career path after graduate school. Thank you to Professor Laura Kubatko for helpful comments as a member of my candidacy and dissertation committees. Thank you to Professor Elizabeth Stasny for her guidance as well.
Thank you to the Faculty Club lunch crew: Casey Davis, John Lewis, and Professor Christopher Hans for the sophisticated and intellectually stimulated conversations over a “Brutus approved” menu. Thanks to fellow graduate students Robert Ashmead, Jiangyong (Matt) Yin, Agniva Som, and Stephen (Bahamamama) Bamattre.

Thank you to both sides of my enormous Catholic family. Unfortunately, listing everyone would take a few pages, so thank you to all the Schneiders and Offenbergers who have supported me over the years. In particular, thank you to my parents Bill and Judy Schneider, my grandparents Herman and Mary Rita Offenberger, and my sister Janell (or Janelle, or maybe even Janel, depending on how she’s spelling it this week) Schneider.

And of course, thank you to my beautiful fiancée, Jingjing Yan for being a great statistician and an even better person.
Vita

March 26, 1986 ............................ Born - Newport, Ohio, USA

2009 ............................. B.S. Honors Accounting

2009 ............................. B.S. Finance

2012 ............................. M.S. Statistics

2009-2010 ........................ Distinguished University Fellow, The Ohio State University.

2010-2011 ........................ Graduate Teaching Assistant, The Ohio State University.

2011-2012 ........................ Graduate Research Assistant, The Ohio State University.

2012-2013 ........................ Graduate Instructor, The Ohio State University.

2013-present ........................ Distinguished University Fellow, The Ohio State University.

Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Outline</td>
<td>5</td>
</tr>
<tr>
<td>2. Background</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Stochastic Differential Equations</td>
<td>6</td>
</tr>
<tr>
<td>2.1.1 The Itô Integral</td>
<td>7</td>
</tr>
<tr>
<td>2.1.2 Itô’s Lemma</td>
<td>13</td>
</tr>
<tr>
<td>2.1.3 Lamperti Transform</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Likelihood Inference for SDEs</td>
<td>15</td>
</tr>
<tr>
<td>2.2.1 Continuous-Time Observations</td>
<td>15</td>
</tr>
<tr>
<td>2.2.2 Discrete-Time Observations</td>
<td>17</td>
</tr>
<tr>
<td>2.2.3 Models with Closed-Form Transition Densities</td>
<td>17</td>
</tr>
<tr>
<td>3. Approximation Methods</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Early Approximations</td>
<td>22</td>
</tr>
<tr>
<td>3.2 Solving the Forward Equation</td>
<td>28</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Number (out of 100000 repetitions) of estimated likelihoods where a negative estimate of the transition density was obtained for at least one $\theta_0$ value using the Hermite approximation [Aït-Sahalia, 2002a] for various time steps $\Delta$. The initial and terminal observations were simulated exactly from the stationary and conditional distribution of (2.19), respectively. The results presented are for true parameter values of $\theta_0 = 2$ and $\theta_1 = -3$.</td>
</tr>
<tr>
<td>4.2</td>
<td>Mean (standard deviation) of the observed proportion (based on 1000 simulated datasets each) of negative estimated likelihoods using the Hermite approximation. The estimated likelihoods were calculated for 100 possible values of $\theta_0 \in [0, 5]$, where the true parameter value of $\theta_0 = 2$ and $\theta_1$ is fixed at the true value -3. These quantities are computed for various time steps $\Delta$ and various numbers of observed data points $N$. The estimated standard errors for these estimates were at most 0.0015.</td>
</tr>
<tr>
<td>4.3</td>
<td>Mean and variance (over $\theta_0$) of the difference between the true log-likelihood and the estimated log-likelihood for $N = 1000$ observations observed $\Delta = 0.1$ apart in time from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ for different numbers of fill-in points between observations $K$ and numbers of Monte Carlo samples $M$.</td>
</tr>
<tr>
<td>4.4</td>
<td>Mean and variance (over $\theta_0$) of the difference between the true and SMC-based estimates of log transition density for extreme value $X(46.5)$ and non-extreme value $X(40)$ for various numbers of subintervals $K$ and $M = 100$ Monte Carlo samples.</td>
</tr>
<tr>
<td>5.1</td>
<td>Two possible realizations of a Latin square over discrete space with $n = 5$ points.</td>
</tr>
</tbody>
</table>
5.2 Two possible realizations (corresponding to the discrete realizations in Table 5.1) of a Latin square over continuous space with \( n = 5 \) points. The sampled points are obtained by uniformly sampling within each of the shaded regions. .................. 93

6.1 A comparison of the true MLE for 1000 simulations of an OU process given by (6.1) with the MLEs obtained using two SMC-based methods: the “Naive” space-filling method and the “SKBO” method. For each SMC-based method we vary \( K \) and \( M \) to control the accuracy of the SMC approximation. We also vary the number of points sampled in the parameter space. The bootstrap standard error for the bias, SD, and RMSE is bounded above by 0.02. .................. 120

6.2 A comparison of the estimated MLEs from 1000 simulations of a GCIR process given by (6.3) with the MLEs obtained using two SMC-based methods: the “Naive” space-filling method and the “SKBO” method. For each SMC-based method we vary \( K \) and \( M \) to control the accuracy of the SMC approximation. We also vary the number of points sampled in the parameter space. The bootstrap standard error for the bias, SD, and RMSE is bounded above by 0.03. .................. 123

6.3 Maximum likelihood estimates and maximized log-likelihoods of the geometric Brownian motion model and estimated MLEs and maximized log-likelihoods of the generalized geometric Brownian motion model for Apple, Inc., Hewlett-Packard Co., and Yahoo! Inc daily adjusted closing stock prices from May 17th, 2004 through May 16th, 2014. . . 128
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Euler approximation (with $\delta = 0.01$) of $dX(t) = \sin(X(t))dt + dW_t$ process observed until time $T = 100$ (top) and $T = 10000$ (bottom). These figures illustrate the behavior of the process used in Beskos and Roberts [2005] to demonstrate Exact Algorithm 1 (EA1).</td>
</tr>
<tr>
<td>4.1</td>
<td>Density plots of realized Monte Carlo samples of the Euler approximation to the Ornstein-Uhlenbeck transition density for various numbers of fill-in points, $K$ (blue lines). The true parameter values for the process are $\theta_0 = 2$ and $\theta_1 = -3$. The black lines represent the true transition density of $X(1)$ given $X_0 = 0.968$, from the conditional distribution of (2.19).</td>
</tr>
<tr>
<td>4.2</td>
<td>Approximate simulated variates based on only $X_0$ (Pedersen) and approximate simulated variates based on both $X_0$ and $X(\Delta)$ (Durham and Gallant) for $\Delta = 1$. The gray lines indicate simulated $X(\tau)$ values and the black lines indicate the calculation of the Monte Carlo variate.</td>
</tr>
<tr>
<td>4.3</td>
<td>True log-likelihoods for $N$ data values simulated from an OU model with $\Delta$ time between observations and true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ (black lines) and corresponding estimated log-likelihoods based on the Hermite expansion of Aït-Sahalia [2002a] (red lines).</td>
</tr>
<tr>
<td>4.4</td>
<td>True log-likelihoods for $N$ data values simulated from an OU model with $\Delta$ time between observations and true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ (black lines) and corresponding estimated log-likelihoods based on the SMC scheme of Pedersen [1995a] with $K = 5$ subintervals and $M = 25$ Monte Carlo samples (red lines).</td>
</tr>
</tbody>
</table>
4.5 Difference between true log-likelihoods and estimated log-likelihoods for $N$ data values simulated from an OU model with $\Delta$ time between observations and true parameter values $\theta_0 = 2$ and $\theta_1 = -3$. The SMC-based estimates are obtained based on the Pedersen’s scheme with $K = 5$ subintervals and $M = 25$ Monte Carlo samples.

4.6 Difference between the true log-likelihood and estimated log-likelihood obtained via SMC for $N = 2$ observations observed $\Delta = 0.1$ apart in time from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ as a function of $\theta_0$ for different numbers of fill-in points between observations $K$ and numbers of Monte Carlo samples $M$. The horizontal red line corresponds to no error.

4.7 Difference between true log-likelihood and estimated log-likelihood obtained via SMC for $N = 1000$ observations observed $\Delta = 0.1$ apart in time from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ as a function of $\theta_0$ for different numbers of fill-in points between observations $K$ and numbers of Monte Carlo samples $M$. The horizontal red line corresponds to no error. We note the very different scales of the left and right columns, which correspond to $M = 100$ and $M = 10000$ Monte Carlo samples, respectively.

4.8 Individual log transition densities and overall log-likelihood of $N = 1000$ observations spaced $\Delta = 0.1$ in time from an OU model with $\theta_0 = 2$ and $\theta_1 = -3$.

4.9 Difference between true and estimated log transition densities obtained via SMC for extreme value $X(46.5)$ and non-extreme value $X(40)$ for various numbers of subintervals $K$ and $M = 100$ Monte Carlo samples. The horizontal red line corresponds to no error. We note the very different scales of the left and right columns, which correspond to the extreme value and non-extreme value, respectively.

4.10 Sampling distribution of log transition density and transition density estimates for extreme value $X(46.5)$ (top) and non-extreme value $X(40)$ (bottom) evaluated at true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ using $K = 50$ subintervals. The black x symbols below the x-axis indicate which values of the likelihood were observed. In the non-extreme case, the observations are less skewed, and these x symbols create a solid black line.
4.11 Smoothed estimates obtained using Pedersen's sampler for various combinations of the number of data points $N$, the number of fill-in points $K$, and the number of Monte Carlo samples $M$. The data were obtained from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ with observations spaced $\Delta = 0.1$ in time. .......................... 86

4.12 Sampling distribution of log transition density estimates for extreme value $X(46.5)$ evaluated at true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ using $K = 50$ subintervals obtained using the sampler of Durham and Gallant (left) and Pedersen’s sampler (right). The top row of histograms are drawn at their respective default scales, while the bottom row is plotted on the same scale to highlight the skew of the distribution of Pedersen's estimates. .......................... 87

5.1 Realization of Latin hypercube used to obtain 10 initial points for SKBO estimation of MLE for OU process with true parameter $\theta_0 = 2$ and $\theta_1 = -3$. .......................... 94

5.2 Initial points (squares), previously added points (gray circles), point added via SKBO on current iteration (red filled circle), estimated MLE on current iteration (dashed line), and true MLE (solid line) for $N = 1000$ observations $\Delta = 0.1$ apart in time simulated from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$. The SMC-based estimates were obtained using $K = 5$ fill-in points and $M = 25$ Monte Carlo samples for each pair of adjacent observations. The SKBO algorithm was stopped when the change to $\hat{\theta}$ was less than 0.01 for five consecutive iterations, which was after 36 points had been added in this illustrative run. .......................... 109

5.3 Approximate confidence regions based on the likelihood ratio test for $N = 1000$ observations from an OU process with $\theta_0 = 2$ and $\theta_1 = -3$ constructed using the true log-likelihood (blue) and kriging mean (red). 112

6.1 (a) A contour plot of the discretized log-likelihood for the OU process given by (6.1) with $\theta_0 = 2$ and $\theta_1 = -3$. The solid horizontal and vertical lines denote the exact MLEs of $\theta_0$ and $\theta_1$ respectively, and the dashed lines denote the SKBO-based estimate. For the SKBO method, the squares indicate the initial parameter values, and the circles denote the values added sequentially. (b) is a zoomed in version of (a). . . . 118
6.2 Daily adjusted closing price for Apple, Inc., Hewlett-Packard Co., and Yahoo! Inc. ........................................... 125

6.3 Normal Q-Q plot of log return for Apple, Inc., Hewlett-Packard Co., and Yahoo! Inc. ........................................... 127
Chapter 1: Introduction

Many phenomena that arise in finance, biology, ecology, and other areas are modeled in continuous time using a real-valued diffusion process, \( \{ X_t \} \), that is the solution to the stochastic differential equation (SDE)

\[
dX(t) = \mu(X(t), \theta) \, dt + \sigma(X(t), \theta) \, dW(t), \quad 0 \leq t \leq T, \tag{1.1}
\]

where \( X_0 = x_0 \) is the initial value of the process and \( \{ W(t) \} \) is a standard Brownian motion. We assume that the drift and the diffusion functions, \( \mu(\cdot, \cdot) \) and \( \sigma(\cdot, \cdot) \) respectively, are known up to the parameter vector \( \theta \in \Theta \), where \( \Theta \) is some compact set in \( \mathbb{R}^p \). We further assume the drift and diffusion functions are locally Lipschitz with linear growth bounds so that a weakly unique solution to (1.1) exists. Suppose that we observe the process \( \{ X(t) \} \) at time points \( t_i \) (\( i = 1, \ldots, N \)) where \( 0 = t_0 < t_1 < \cdots < t_N \), and let \( X = (X(t_1), \ldots, X(t_N))^T \). In this dissertation we are interested in maximum likelihood estimation of \( \theta \) and associated confidence bounds based on the data \( X \).

Let \( p(x|x(t_{i-1}), \theta) \) represent the conditional probability density of \( X(t_i) \) given \( X(t_{i-1}) = x(t_{i-1}) \) evaluated at \( x \) for a given set of parameters \( \theta \in \Theta \). Treating \( X(0) = X_0 \) as fixed, we can then use the Markov property to write the likelihood of
the data as the product of these individual transition densities

\[ L(\theta|X) = \prod_{i=1}^{N} p(X(t_i)|X(t_{i-1}), \theta). \]  

(1.2)

When the transition density is known, likelihood calculation and its maximization with respect to \( \theta \in \Theta \) for a given set of data discretely observed from (1.1) is straightforward. As the transition density does not exist in closed-form except for a handful of cases, approximations are typically necessary; see for example Hurn et al. [2007] for a recent overview. These methods are often separated into four groups: (1) approximations derived by numerically solving the Kolmogorov forward equation [Lo, 1988], (2) closed-form Hermite expansions of the transition density [Aït-Sahalia, 2002a, 2008], (3) sequential Monte Carlo (SMC) [Pedersen, 1995a, Santa-Clara, 1997, Elerian et al., 2001, Brandt and Santa-Clara, 2002b, Durham and Gallant, 2002, Lin et al., 2010], and (4) methods based on the exact simulation of diffusions [Beskos et al., 2006b].

After illustrating the details of these four methods of approximation, we compare and contrast them, first in terms of approximating the transition density for a pair of observations. Past comparisons of the performance of some of these methods in estimating the transition density have been made by Aït-Sahalia [2002a,b] and Stramer and Yan [2007a], among others. We summarize these results and supplement them with additional guidance based on simulation studies. We then turn to investigating the performance of the methods in terms of estimating the log-likelihood, the sum of the individual log transition densities for each sequential pair of points. Presumably because of this straightforward construction from the log transition densities, the comparison based on the estimation of the log-likelihood has received less attention in the past. However this is an important concern in practice, as likelihood inference
is often of interest in increasingly large datasets. We find that each of the methods encounters difficulty with estimating the transition density for those observations which are in the tail of the conditional distribution given the previous observation. Although these problems may be brushed aside as unlikely when considering only the transition density, as the number of observed data values, $N$, grows, extreme observations will be encountered more and more often. We then use simulated data to demonstrate that this problem may be solved for the SMC-based likelihood estimation procedure by simply increasing the number of Monte Carlo samples $M$ relative to the number of simulated fill-in values $K$. While this solution incurs additional computational cost, it is a much simpler fix than those necessary to remedy the problems encountered by the Hermite expansion. With this mind, we then provide guidance on choosing $K$ and $M$, the user-specified inputs to SMC, which includes considering the number of observed data points $N$ to supplement previously established rules.

After motivating the choice of SMC by the relative ease of estimating the log-likelihood, we focus on this procedure, which can be made arbitrarily accurate at the expense of more computation. One of our goals is to ease this computational burden associated with repeatedly obtaining Monte Carlo estimates of the log-likelihood over the parameter space. Much of the previous work has focused on efficiently estimating the likelihood at a fixed $\theta$. An approximate MLE can then be obtained by maximizing this estimated likelihood over $\Theta$. In some cases, the derivatives of the log-likelihood with respect to $\theta$ can be obtained from the simulated values used to produce the estimate of the likelihood as in Stramer and Yan [2007a] and gradient ascent optimization is straightforward. Typically, however, these derivatives must be obtained numerically, which adds a significant computational burden. Although the
underlying log-likelihood may be smooth as a function of $\theta$, the Monte Carlo estimates will be subject to variability and thus will be much less amenable to derivative calculation.

We propose an efficient Gaussian-process-based method for exploring $\Theta$ which accounts for the inherent Monte Carlo variability of the simulated likelihood method and does not require knowledge of the gradient of the log-likelihood. Our sequential method offers significant computational efficiency over a naive approach based on estimating the likelihood over a grid of possible parameter values. Estimating the likelihood over a grid can be computationally demanding when $\theta$ is of low dimension, and practically impossible as the dimension grows to be moderate or large.

By using a global search criterion borrowed from the computer experiments literature called the “expected improvement” [Jones et al., 1998, Schonlau et al., 1998, Williams et al., 2000], we alleviate difficulties with local maxima that may be encountered by gradient ascent methods. This criterion can be thought of as balancing the uncertainty in estimating the discretized log-likelihood at unexplored parameter values with the desire to find parameter values near the current maximum that have a higher log-likelihood.

By maximizing this criterion conditional on the Gaussian process model fit to the current set of observations, we obtain a sequential sample of parameter values, thus avoiding the need to use a regular grid for likelihood evaluation. This model fit and subsequent maximization relies on an interpolation technique from the geostatistics literature known as kriging [Matheron, 1963], which predicts the value of a function at a point using an average of known values nearby, weighted according to their respective spatial covariances with the point of interest. The kriging mean
provides a useful estimate of the log-likelihood at unobserved $\theta$ values. This kriging approach allows for estimation of the MLE of $\theta$ and straightforward quantification of its uncertainty. We demonstrate the flexibility and usefulness of this approach using simulations studies and by considering historical stock market data.

1.1 Outline

This thesis is organized as follows. Chapter 2 provides the necessary background information for defining and understanding the statistical properties of SDEs. Chapter 3 discusses four groups of methods used to approximate the transition density and thus the likelihood function of SDEs. Chapter 4 compares these methods in terms of the accuracy of the approximation and the computational efficiency of each method. Chapter 5 illustrates the SKBO search strategy for efficiently estimating the MLE of $\theta$ based on SMC output. Chapter 6 demonstrates the usefulness and flexibility of SKBO on real data by fitting a variety of models to the stock prices of various companies. Chapter 7 discusses conclusions, extensions, and future work.
Chapter 2: Background

In this chapter we review the mathematical foundations of SDEs and then briefly discuss likelihood inference based on discrete realizations of an SDE.

2.1 Stochastic Differential Equations

We begin with the foundations of stochastic calculus, the standard Brownian motion (or Wiener process), first described in Brown [1828]. Brownian motion was first used to model the motion of pollen particles throughout a fluid and was used in Einstein [1905] to prove the existence of atoms and molecules. Formally, a standard Brownian motion is a stochastic process \( \{ W(t), t \geq 0 \} \) defined on a probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \) with the following properties:

1. \( W(0) = 0 \) a.s..

2. The increments, \( W(t) - W(s) \), follow a normal distribution with mean zero and variance \( t - s \) for each \( 0 \leq s < t \).

3. \( \{ W(t), t \geq 0 \} \) has independent increments. That is, if \( 0 < t_1 < \cdots < t_n \), then \( W(t_1), W(t_2) - W(t_1), \ldots, W(t_n) - W(t_{n-1}) \) are independent random variates.
4. \((W(t_1), W(t_2) - W(t_1), \ldots, W(t_n) - W(t_{n-1}))\) follow a multivariate normal distribution with mean \(0\) and the diagonal covariance matrix which has element \(i, i\) given by \(t_i - t_{i-1}\) for \(i = 1, \ldots, n\).

It is also known that the paths of a standard Brownian motion are continuous but nowhere differentiable [see, for example, Shreve, 2004].

**Brownian Bridge** A Brownian bridge can be thought of as a Brownian motion conditioned on a specific endpoint. That is, the process the standard Brownian bridge \(B(t)\) is defined as

\[
B(t) = (W(t)|W(1) = 0), \quad 0 \leq t \leq 1.
\]  

(2.1)

A Brownian bridge with endpoints \(a\) and \(b\) over the interval \(t_1 \leq t \leq t_2\) may be obtained by shifting and scaling \(B(t)\) appropriately [Shreve, 2004].

### 2.1.1 The Itō Integral

Consider a *simple* function \(h(t)\) for \(t \geq 0\) (a piecewise constant function in \(t\)). For illustration purposes, say that \(h(t)\) represents the number of shares of stock held at time \(t\). Now, assuming for the time being that it is reasonable to represent the price of a stock using a Brownian motion process, we can model the gain (or loss) of the portfolio by finding the gain in each transaction (the change in \(h(t)\)) made at times \(0 = t_0 < t_1 < \cdots < t_n = t\). That is, we begin by holding \(h(t_0)\) shares until some time \(t_1\), when we switch to holding \(h(t_1)\) shares until some time \(t_2\), and then we switch to holding \(h(t_2)\) shares, etc. We can compute the gain of each of these transactions as \(h(t_0) (W(t_1) - W(t_0))\), \(h(t_1) (W(t_2) - W(t_1))\), and so on, the number of shares held over the time period multiplied by the change in price. Generalizing
this for an arbitrary time \( t \in [t_{i-1}, t_i) \), the cumulative gain up to time \( t \) is

\[
I(t, h) = \sum_{j=1}^{i-1} h(t_j) [W(t_j) - W(t_{j-1})] + h(t_{i-1}) [W(t) - W(t_{i-1})].
\]  

(2.2)

Expression (2.2) is known as the \( \text{Itô integral} \) [\text{Itô, 1944}] of the simple function \( h(t) \) with respect to \( W(t) \).

This simple idea can be generalized to more complicated functions \( h(\cdot) \) in the following way. Let \( h_n(\cdot) \) be a sequence of simple functions such that

\[
\lim_{n \to \infty} E \int_0^T \|h_n(t) - h(t)\|^2 dt = 0.
\]

This sequence will exist as long as \( h(t) \) is bounded and has continuous sample paths. Throughout this dissertation we make the assumption that all the expectations exist and are finite. After choosing such a sequence, define the \( \text{Itô integral} \) for a general function \( h(t) \) with respect to the standard Brownian motion \( W(t) \) to be

\[
I(t, h) = \int_0^t h(s)dW(s) = \lim_{n \to \infty} \int_0^t h_n(u)dW(u), \quad 0 \leq t \leq T.
\]  

(2.3)

Observe that \( \{I(t, h), t \geq 0\} \) is a stochastic process with the following properties:

- \( I(0, h) = 0 \) a.s. and \( \{I(t, h), t \geq 0\} \) has continuous sample paths.

- \( I(t, h) \) is a martingale. That is, for \( s < t \),

\[
E[I(t, h)|I(s, h)] = I(s, h).
\]

- \( I(t, h) \) satisfies the Markov property. That is, for \( s_1 < s_2 < t \) the conditional distribution of \( I(t, h) \) given \( I(s_1, h) \) and \( I(s_2, h) \) does not depend on \( I(s_1, h) \).

- (Linearity) \( I(t, \alpha h + \beta g) = \alpha I(t, h) + \beta I(t, g) \).

- The random variable \( I(t, h) \) is normally distributed with mean 0 and variance \( \int_0^t h^2(s)ds \).
For a proof of these properties, see Chapter 4 of Shreve [2004].

With the help of the Itô integral, one can define an Itô process which takes the form

\[ X(t) = X_0 + \int_0^t \mu(X(u), \theta) \, du + \int_0^t \sigma(X(u), \theta) \, dW(u), \quad 0 \leq t \leq T, \tag{2.4} \]

where \( W(t) \) is a standard Brownian motion. Henceforth we assume that the functions \( \mu(\cdot, \cdot) \) and \( \sigma(\cdot, \cdot) \) are such that the integrals above exist. In expression (2.4) the first integral is the usual Riemann integral and the second is an Itô integral as defined in (2.3). The functions \( \mu(\cdot, \cdot) \) and \( \sigma(\cdot, \cdot) \) are known as the drift and the diffusion, respectively, and are assumed to be known up to the \( p \)-dimensional parameter vector \( \theta \in \Theta \), where \( \Theta \) here is assumed to be some compact set in \( \mathbb{R}^p \).

Equation (2.4) is often expressed in the more notationally convenient differential form of

\[ dX(t) = \mu(X(t), \theta) \, dt + \sigma(X(t), \theta) \, dW(t), \quad 0 \leq t \leq T, \tag{2.5} \]

where \( X(0) = X_0 \) is the initial value of the process.

Existence of a process satisfying (2.4) is not a trivial question. It can be shown that under certain conditions, an Itô process does exist. We now summarize such conditions.

A strong solution to (2.4) (or equivalently (2.5)) will exist if both the drift and diffusion are Lipschitz continuous. That is, there exists a positive constant \( C_0 \), such that for all \( \theta, x, \) and \( y \)

\[ |\mu(x, \theta) - \mu(y, \theta)| + |\sigma(x, \theta) - \sigma(y, \theta)| \leq C_0 |x - y|. \]
The solution will be non-explosive if \( E(X(0)^2) < \infty \) and if there exists a positive constant \( C_1 \), such that for all \( \theta \) and \( x \)
\[
|\mu(x, \theta)|^2 + |\sigma(x, \theta)|^2 \leq C_1 \left( 1 + x^2 \right).
\]
A strong solution is rarely available in closed-form, however numerical methods have been developed to approximate the solution.

Under weaker assumptions, a weak solution to (2.4) exists. While existence of a strong solution implies existence of a weak solution, the converse is not generally true. A famous example of an SDE which has a weak solution but not a strong solution is given by the solution to Tanaka’s equation
\[
dX(t) = \text{sgn}(X(t))dW(t), \quad 0 \leq t \leq T,
\]
where \( X(0) = X_0 \) is the initial value of the process and \( \text{sgn}(x) \) is the sign function which is equal to -1 when \( x < 0 \) and equal to 1 when \( x \geq 0 \). Note that \( \text{sgn}(x) \) does not satisfy the Lipschitz continuity required for a strong solution to exist. It can be shown [see Øksendal, 2003] that a weak solution exists for \( X(t) \) satisfying
\[
dX(t) = \sigma dW(t), \quad 0 \leq t \leq T,
\]
where \( X(0) = 0 \) if \( \sigma^2 = 1 \). Since \( [\text{sgn}(x)]^2 = 1 \), a weak solution exists. In particular, the solution \( X(t) \) follows the same distribution as a Brownian motion. While a strong solution implies that, given \( X(t) \), the particular Brownian motion can be determined, in this case, we cannot determine which Brownian motion generated \( X(t) \), as any Brownian motion could have generated \( X(t) \). This is due to the reflection principle of the Brownian motion.
Quadratic Variation

Let \( \{X(t), t \geq 0\} \) be a stochastic process and let \( \pi \) be the partition

\[
0 = t_0 < t_1 < \cdots < t_n = T.
\]

The norm of the partition \( \pi \) is given by \( |\pi| = \max_{i \in 1, \ldots, n} (t_i - t_{i-1}) \). We define the quadratic variation of the process \( \{X(t), 0 \leq t \leq T\} \) to be

\[
[X, X](T) = \lim_{|\pi| \to 0} \sum_{i=1}^{n} [X(t_i) - X(t_{i-1})]^2.
\]

**Lemma 2.1.1.** If \( \{W(t), 0 \leq t \leq T\} \) is a standard Brownian motion then the quadratic variation of \( W(t) \) up to time \( t \) is \( [W, W](t) = t \) a.s.

**Proof.** Note that since the increments of Brownian motion have mean zero,

\[
\lim_{|\pi| \to 0} \mathbb{E} \left[ \sum_{i=1}^{n} (W(t_i) - W(t_{i-1}))^2 \right] = \lim_{|\pi| \to 0} \sum_{i=1}^{n} \mathbb{E} \left[ (W(t_i) - W(t_{i-1}))^2 \right]
\]

\[
= \lim_{|\pi| \to 0} \sum_{i=1}^{n} \text{Var}(W(t_i) - W(t_{i-1}))
\]

\[
= \lim_{|\pi| \to 0} \sum_{i=1}^{n} (t_i - t_{i-1})
\]

\[
= t. \quad (2.7)
\]

In addition,

\[
\lim_{|\pi| \to 0} \text{Var} \left[ \sum_{i=1}^{n} (W(t_i) - W(t_{i-1}))^2 \right]
\]

\[
= \lim_{|\pi| \to 0} \sum_{i=1}^{n} \text{Var} \left[ (W(t_i) - W(t_{i-1}))^2 \right] \quad \text{(independent increments)}
\]

\[
= \lim_{|\pi| \to 0} \sum_{i=1}^{n} \mathbb{E} \left\{ \left[ (W(t_i) - W(t_{i-1}))^2 - (t_i - t_{i-1}) \right]^2 \right\}
\]

\[
= \lim_{|\pi| \to 0} \sum_{i=1}^{n} \left\{ \mathbb{E} \left[ (W(t_i) - W(t_{i-1}))^4 \right] - 2(t_i - t_{i-1}) \mathbb{E} \left[ (W(t_i) - W(t_{i-1}))^2 \right] \right\}
\]
\[
\left\{ \begin{align*}
&+ (t_i - t_{i-1})^2 \\
&= \lim_{|\pi| \to 0} \sum_{i=1}^{n} \left\{ 3(t_i - t_{i-1})^2 - 2(t_i - t_{i-1})^2 + (t_i - t_{i-1})^2 \right\} \\
&= \lim_{|\pi| \to 0} \sum_{i=1}^{n} 2(t_i - t_{i-1})^2 \\
&\leq \lim_{|\pi| \to 0} \sum_{i=1}^{n} 2|\pi|(t_i - t_{i-1}) = \lim_{|\pi| \to 0} 2|\pi|t = 0,
\end{align*} \right.
\]

(2.8)

using the definition of $|\pi|$ for a bound.

We have shown that $\sum_{i=1}^{n} (W(t_i) - W(t_{i-1}))^2$ converges to $t$ in $L^2$ as the norm of the partition, $|\pi| \to 0$. Therefore there must exist some subsequence of $|\pi|$ such that the quadratic variation converges to $t$ almost surely. It can be shown [see Revuz and Yor, 1999] that when the subsequence is chosen so that the intervals of $\pi$ are nested, the convergence is almost sure. This result is often represented informally as $dW(t)dW(t) = dt$.

\[\text{Lemma 2.1.2. If } \{W(t), 0 \leq t \leq T\} \text{ is a standard Brownian motion then } dW(t)dt = dt dt = 0.\]

\[\text{Proof. Observe that}\]

\[
\lim_{|\pi| \to 0} \sum_{i=1}^{n} (t_i - t_{i-1})^2 \leq |\pi| \sum_{i=1}^{n} (t_i - t_{i-1}) = \lim_{|\pi| \to 0} |\pi| t = 0.
\]

(2.9)

Also,

\[
\lim_{|\pi| \to 0} \sum_{i=1}^{n} (W(t_i) - W(t_{i-1}))(t_i - t_{i-1}) \leq \lim_{|\pi| \to 0} \left( \max_{1 \leq k \leq n} |W(t_k) - W(t_{k-1})| \sum_{i=1}^{n} (t_i - t_{i-1}) \right) \\
= \lim_{|\pi| \to 0} \left( \max_{1 \leq k \leq n} |W(t_k) - W(t_{k-1})| t \right) = 0,
\]

(2.10)

where (2.10) follows from the continuity of $W(t)$. \qed
The previous results allow us to easily determine the quadratic variation of the Itô process $X(t)$ defined in (2.4) as well. Using the stochastic rules described above we find:

\begin{align*}
dX(t)dX(t) &= (\mu(X(t), \theta) dt + \sigma(X(t), \theta) dW(t)) (\mu(X(t), \theta) dt + \sigma(X(t), \theta) dW(t)) \\
&= (\mu^2(X(t), \theta) dt^2) + 2 (\mu(X(t), \theta) \sigma(X(t), \theta) dt dW(t)) \\
&+ (\sigma^2(X(t), \theta) dW(t)dW(t)) \\
&= \sigma^2(X(t), \theta) dt, \quad (2.11)
\end{align*}

since the $dt dt$ and $dW(t)dt$ terms vanish.

### 2.1.2 Itô’s Lemma

One of the most important results of stochastic calculus is given by Itô’s lemma, which states that for a process given by (2.5) and any twice differentiable function $f(t, x)$, we have

\begin{align*}
df(t, X(t)) &= \left[ f_t(t, X(t)) + \mu(X(t), \theta) f_x(t, X(t)) + \frac{1}{2} \sigma^2(X(t), \theta) f_{xx}(t, X(t)) \right] dt \\
&+ \sigma(X(t), \theta) f_x(t, X(t)) dW(t), \quad (2.12)
\end{align*}

where

\begin{align*}
f_t(t, X(t)) &= \left. \frac{d}{dt} f(t, x) \right|_{x=X(t)}, \\
f_x(t, X(t)) &= \left. \frac{d}{dx} f(t, x) \right|_{x=X(t)}, \\
and f_{xx}(t, X(t)) &= \left. \frac{d^2}{dx^2} f(t, x) \right|_{x=X(t)}.
\end{align*}

The result in (2.12) has many applications. In particular it can aid with finding closed-form solutions for a handful of processes. We illustrate this idea in later sections.
2.1.3 Lamperti Transform

In general, the diffusion term of (2.5) tends to cause more difficulties in practical applications than the drift term. For a one-dimensional SDE \( \{X(t), 0 \leq t \leq T\} \), the \textit{Lamperti transform} can be used to obtain a unit diffusion with \( \sigma(X(t), \theta) = 1 \).

This transformation is required by some estimation schemes that we will encounter later and often improves performance of the estimators even in cases where it is not required. This transformation is given by

\[
Y(t) \equiv \gamma(X(t); \theta) = \int_c^{X(t)} \frac{du}{\sigma(u; \theta)},
\]

for any \( c \) in the domain of \( X(t) \). Using Itô’s Lemma,

\[
dY(t) = \frac{\partial}{\partial t} \gamma(X(t); \theta) dt + \frac{\partial}{\partial x} \gamma(X(t); \theta) dX(t) + \frac{1}{2} \frac{\partial^2}{\partial x^2} \gamma(X(t); \theta) dX(t) dX(t) \\
= 0 \cdot dt + \frac{1}{\sigma(X(t); \theta)} [\mu(X(t), \theta) dt + \sigma(X(t), \theta) dW(t)] \\
\quad + \frac{\partial \sigma}{\partial x}(X(t); \theta) \sigma^2(X(t); \theta) dt \\
= \frac{1}{\sigma(X(t); \theta)} \mu(X(t), \theta) dt + dW(t) + \frac{1}{2} \frac{\partial \sigma}{\partial x}(X(t); \theta) dt \\
= \frac{1}{\sigma(\gamma^{-1}(Y(t); \theta); \theta)} \mu(\gamma^{-1}(Y(t); \theta), \theta) dt + \frac{1}{2} \frac{\partial \sigma}{\partial x}(\gamma^{-1}(Y(t); \theta); \theta) dt + dW(t) \\
= \tilde{\mu}(Y(t); \theta) dt + dW(t),
\]

where

\[
\tilde{\mu}(y; \theta) = \frac{\mu(\gamma^{-1}(y; \theta); \theta)}{\sigma(\gamma^{-1}(y; \theta); \theta)} - \frac{1}{2} \frac{\partial \sigma}{\partial x}(\gamma^{-1}(y; \theta); \theta),
\]

is the reparametrized drift function that depends on the original drift and diffusion functions and the inverse of the Lamperti transform function \( \gamma^{-1}(\cdot, \cdot) \). Note that the Lamperti transform may not exist for a given multi-dimensional diffusion. Conditions
where the Lamperti transform does exist for a multi-dimensional diffusion are given in Aït-Sahalia [2008].

2.2 Likelihood Inference for SDEs

In the previous sections, we have focused on the mathematical or stochastic properties of SDEs. We now turn to the statistical problem of estimating the parameter \( \theta \) given observations from a stochastic differential equation. We distinguish between two cases. First we describe the situation where we have access to continuous-time observations. Secondly we consider the case when observations are only available discretely in time.

2.2.1 Continuous-Time Observations

Given an observed path \( \{X(t), 0 \leq t \leq T\} \), we start our inference exposition with a formal definition of the likelihood function. In general, to compare two distributions, say \( P_1 \) and \( P_2 \), with densities \( f_1 \) and \( f_2 \) with respect to some dominating measure \( \lambda \), we can use the Radon-Nikodym derivative given by

\[
\frac{dP_1}{dP_2}(x) = \frac{f_1(x)}{f_2(x)}.
\]  

(2.14)

For a given \( x \), high values of this ratio indicate that it is more likely that \( x \) is a draw from the distribution \( P_1 \) rather than from \( P_2 \). When \( f_1 \) and \( f_2 \) are specified through a parameter \( \theta \), (2.14) is called the likelihood function.

Similarly, in the SDE context, let \( \{X(t), 0 \leq t \leq T\} \) be an Itô process described via

\[
dX(t) = \mu(X(t), \theta) \, dt + \sigma(X(t), \theta) \, dW(t)
\]
and define the driftless process \( \{M(t), 0 \leq t \leq T\} \) described via
\[
dM(t) = \sigma(X(t), \theta) dW(t).
\]

Let \( \mathbb{P} \) and \( \mathbb{M} \) be the distributions induced by \( \{X(t), 0 \leq t \leq T\} \) and \( \{M(t), 0 \leq t \leq T\} \), respectively. Then the Radon-Nikodym derivative is given by Girsanov’s formula [Girsanov, 1960]:
\[
\frac{d\mathbb{P}}{d\mathbb{M}}(X) \equiv G(X; \mu; \sigma) = \exp \left( \int_0^T \frac{\mu(X(s), \theta)}{\sigma^2(X(t), \theta)} dX(s) - \frac{1}{2} \int_0^T \frac{\mu^2(X(s), \theta)}{\sigma^2(X(t), \theta)} ds \right). \tag{2.15}
\]

If \( \mu(\cdot) \) is everywhere differentiable, equation (2.15) can be rewritten as [see, for example, Shreve, 2004]:
\[
G(X; \mu; \sigma) = \exp \left( \int_{X(0)}^{X(t)} \frac{\mu(\nu, \theta)}{\sigma^2(X(t), \theta)} d\nu - \frac{1}{2} \int_0^T \left[ \frac{d}{dX(s)} \frac{\mu(X(s), \theta)}{\sigma^2(X(t), \theta)} \right] ds \\
- \frac{1}{2} \int_0^T \frac{\mu(X(s), \theta)^2}{\sigma^2(X(t), \theta)} ds \right),
\]
where we note that all of the integrals appearing above are Riemann integrals. To compare two models with the same diffusion term but different drift terms, and respective probability measures \( \mathbb{P}_1 \) and \( \mathbb{P}_2 \), we can use (2.15) to compute
\[
\frac{d\mathbb{P}_1}{d\mathbb{P}_2} = \frac{d\mathbb{P}_1}{d\mathbb{M}} \frac{d\mathbb{M}}{d\mathbb{P}_2}. \tag{2.16}
\]

When we have access to a continuous path, inference is fairly straightforward, as we can calculate the elements in \( \theta \) which appear in the expression of \( \sigma(\cdot, \theta) \) as explained in Le Breton [1977] by
\[
\sigma^2(X(t), \theta) = \frac{d[X, X](t)}{dt} = \lim_{n \to \infty} 2^n (X(t) - X(t - 2^{-n}))^2.
\]
Once these elements of $\theta$ have been determined, one can maximize Girsanov’s formula to find the remaining parameters. This procedure is explained in detail in Chapter 6 of Fuchs [2013]. However, as it is impossible to observe data continuously in practice, we now turn to the issues that arise due to discretely-observed data.

2.2.2 Discrete-Time Observations

To examine a more practical setting, suppose that we observe the process $\{X(t), 0 \leq t \leq T\}$ at time points $t_i \ (i = 1, \ldots, N)$ where $0 = t_0 < t_1 < \cdots < t_N$, and let $X = (X(t_1), \ldots, X(t_N))^T$. Let $p(x|x(t_{i-1}), \theta)$ represent the conditional probability density of $X(t_i)$ given $X(t_{i-1}) = x(t_{i-1})$ evaluated at $x$ for a given set of parameters $\theta \in \Theta$. Treating $X(0) = X_0$ as fixed, we can then use the Markov property of the SDE (2.3) to write the likelihood of the data as the product of these individual transition densities

$$L(\theta|X) = \prod_{i=1}^{N} p(X(t_i)|X(t_{i-1}), \theta). \quad (2.17)$$

When the transition density is known, likelihood calculation and its maximization with respect to $\theta \in \Theta$ for a given set of data discretely observed from (2.5) is straightforward. The transition density does not exist in closed-form except for a handful of models. Outside of the three cases that follow, approximations to the transition density are typically necessary.

2.2.3 Models with Closed-Form Transtion Densities

Without loss of generality, assume that we are interested in the transition density from $X(0) = X_0$ to $X(\Delta)$. 

17
Geometric Brownian Motion

The geometric Brownian motion is defined by

\[ dX(t) = \theta_0 X(t) \, dt + \gamma X(t) \, dW(t), \quad 0 \leq t \leq T, \]  

(2.18)

where \( X(0) = X_0 \) is the initial value of the process \( \theta_0 \in \mathbb{R}, \gamma > 0 \), and \( W(t) \) is a standard Brownian motion. Stock prices were modeled by (2.18) in the famous Black-Scholes model [Black and Scholes, 1973] for option pricing and (2.18) is referred as “the model for stock prices” in Hull [2012], a popular introductory finance text.

To determine the solution to (2.18) we use Itô’s Lemma with \( f(t, X(t)) = \log(X(t)) \).

We then compute

\[ f_t(t, X(t)) = 0, \]
\[ f_x(t, X(t)) = \frac{1}{X(t)}, \]
\[ \text{with } f_{xx}(t, X(t)) = -\frac{1}{X(t)^2}, \]

and plug them into (2.12) to get

\[
\begin{aligned}
  d\log(X(t)) & = \left( 0 + \theta_0 X(t) - \frac{1}{2} \gamma^2 X(t)^2 \right) dt + \gamma X(t) \frac{1}{X(t)} dW(t) \\
  & = \left( \theta_0 - \frac{\gamma^2}{2} \right) dt + \gamma dW(t), \quad 0 \leq t \leq T.
\end{aligned}
\]

By the Gaussianity of the Itô integral, we have that conditionally on \( \log(X(0)) \), \( \log(X(\Delta)) \) follows a normal distribution with mean and variance given by

\[
\begin{aligned}
  \mathbb{E}(\log(X(\Delta))|\log(X(0))) & = \log(X(0)) + \left( \theta_0 - \frac{\gamma^2}{2} \right) \Delta \quad \text{and} \\
  \text{Var}(\log(X(\Delta))|\log(X(0))) & = \gamma^2 \Delta.
\end{aligned}
\]

Equivalently, conditionally on \( X(0) \), for \( \Delta > 0 \), \( X(\Delta) \) follows a log-normal distribution with parameters determined from the display above.
Ornstein-Uhlenbeck Process

The *Ornstein-Uhlenbeck* (OU) model [Uhlenbeck and Ornstein, 1930] is defined by

\[
dX(t) = (\theta_0 + \theta_1 X(t)) \, dt + \gamma \, dW(t), \quad 0 \leq t \leq T, \tag{2.19}
\]

where \(X(0) = X_0\) is the initial value of the process, \(\theta_0 \in \mathbb{R}, \theta_1 < 0, \gamma > 0,\) and \(W(t)\) is a standard Brownian motion.

To determine the solution to (2.19) we use Itô’s Lemma with \(f(t, X(t)) = X(t)e^{-\theta_1 t}\).

We then compute

\[
f_t(t, X(t)) = -\theta_1 X(t)e^{-\theta_1 t},
\]

\[
f_x(t, X(t)) = e^{-\theta_1 t},
\]

with \(f_{xx}(t, X(t)) = 0\),

and plug them into (2.12) to get

\[
d \left( X(t)e^{-\theta_1 t} \right) = \left[ -\theta_1 X(t)e^{-\theta_1 t} + (\theta_0 + \theta_1 X(t)) e^{-\theta_1 t} + 0 \right] \, dt + \gamma e^{-\theta_1 t} dW(t)
\]

\[
= \theta_0 e^{-\theta_1 t} \, dt + \gamma e^{-\theta_1 t} dW(t)
\]

for \(0 \leq t \leq T\). The above expression can be written in integral form as

\[
X(t)e^{-\theta_1 t} = X_0 + \int_0^t \theta_0 e^{-\theta_1 s} \, ds + \int_0^t \gamma e^{-\theta_1 s} \, dW_s
\]

\[
= X_0 - \frac{\theta_0}{\theta_1} (e^{-\theta_1 t} - 1) + \int_0^t \gamma e^{-\theta_1 s} \, dW_s.
\]

We then have that

\[
X(t) = X_0 e^{\theta_1 t} - \frac{\theta_0}{\theta_1} (1 - e^{\theta_1 t}) + \int_0^t \gamma e^{-\theta_1 (s-t)} \, dW_s. \tag{2.20}
\]
From the definition of the Itô integral (2.3), we have that the last term of (2.20),
\[
\int_0^t \gamma e^{-\theta_1 (s-t)} dW_s \sim N \left( 0, \int_0^t \left[ \gamma e^{-\theta_1 (s-t)} \right]^2 ds \right)
\]
\[
= N \left( 0, \gamma^2 e^{2\theta_1 t} \int_0^t e^{-2\theta_1 s} ds \right)
\]
\[
= N \left( 0, \gamma^2 e^{2\theta_1 t} \frac{1}{2\theta_1} \left[ 1 - e^{-2\theta_1 t} \right] \right)
\]
\[
= N \left( 0, \frac{\gamma^2}{2\theta_1} \left[ e^{2\theta_1 t} - 1 \right] \right).
\]
Since the first two terms on the right side of (2.20) are constants we have that conditionally on \(X(0) = X_0\), for \(\Delta > 0\), \(X(\Delta)\) is normally distributed with mean and variance given by
\[
\mathbb{E}(X(\Delta)|X_0) = X_0 e^{\theta_1 \Delta} - \frac{\theta_0}{\theta_1} (1 - e^{\theta_1 \Delta}) \quad \text{and}
\]
\[
\text{Var}(X(\Delta)|X_0) = \frac{\gamma^2}{2\theta_1} (e^{2\theta_1 \Delta} - 1).
\]
One can also verify that the stationary distribution for \(\{X(t), t \geq 0\}\) is Gaussian with mean \(-\theta_0/\theta_1\) and variance \(-\gamma^2/(2\theta_1)\).

The OU process has the desirable property that it is mean reverting. That is, when \(X(t)\) is larger than the stationary mean \(-\theta_0/\theta_1\), the drift term \((\theta_0 + \theta_1 X(t))\) is negative, and when \(X(t)\) is smaller than the stationary mean \(-\theta_0/\theta_1\), the drift term \((\theta_0 + \theta_1 X(t))\) is positive. However, it also has the property that \(X(t)\) can be negative, which is undesirable in many financial settings, for example an interest rate or a stock price. The OU process, which can be thought of as the continuous time analogue of the discrete-time autoregressive process of order 1, is commonly used to model currency exchange rates [e.g., Elliott et al., 2005], has applications in physics [e.g., Bibbona et al., 2008], and is used in biology [e.g., Butler and King, 2004].
Cox-Ingersoll-Ross Process

The Cox-Ingersoll-Ross (CIR) model [Cox et al., 1985] is defined by

\[ dX(t) = (\theta_0 + \theta_1)X(t)\,dt + \gamma \sqrt{X(t)}\,dW(t), \quad 0 \leq t \leq T, \]  

(2.21)

where \( X(0) = X_0 \) is the initial value of the process, \( \theta_0 \in \mathbb{R}, \theta_1 < 0, \gamma > 0, \) and \( W(t) \) is a standard Brownian motion. It can be shown [see Feller, 1951, Dyrting, 2004] that under this model, given \( X(0) = X_0 \), for \( \Delta > 0 \),

\[ -\frac{4\theta_1}{\gamma^2(1 - e^{\theta_1\Delta})} X(\Delta) \sim \chi^2_{p,\lambda}, \]

where \( \chi^2_{p,\lambda} \) represents the non-central chi-square distribution with \( p \) degrees of freedom and non-centrality parameter \( \lambda \). The degrees of freedom and non-centrality parameter are given by, respectively

\[ p = \frac{4\theta_0}{\gamma^2} \quad \text{and} \quad \lambda = -\frac{4\theta_1}{\gamma^2(1 - e^{-\theta_1\Delta})} X_0 e^{\theta_1\Delta}. \]

The CIR process shares the mean reversion property of the OU process, but also is restricted to live on the positive real line. The model’s ability to model interest rates has been used in Maghsoodi [1996] to value bonds and has been generalized to allow for stochastic mean and stochastic volatility in Chen [1996].

The need to approximate the transition density

When the process chosen is not one of those discussed above, the transition density \( p(X(\Delta)|X_0; \theta) \) will typically not exist in closed-form and must be approximated. In the next chapter, we will review several approximations to the transition density and evaluate their performance in terms of estimating the transition density. We will also consider the how these approximations extend to estimating the overall log-likelihood and the effect this extension has on performance.
Chapter 3: Approximation Methods

When the process chosen is not one of those discussed in the previous chapter, the conditional distribution of $X(\Delta)$ given $X_0$ for $\Delta > 0$, the transition density $p(X(\Delta)|X_0; \theta)$, will typically not exist in closed-form and must be approximated. We begin with a review of early work in this area.

3.1 Early Approximations

First-Order Approximation

A simple and intuitive method of approximating $p(X(\Delta)|X_0; \theta)$ is based on the Euler-Maruyama approximation of the process. This method approximates the drift and diffusion terms by constant functions over the interval $[0, \Delta)$. That is,

$$
\int_0^\Delta \mu(X(s), \theta) \, ds \approx \int_0^\Delta \mu(X_0, \theta) \, ds = \mu(X_0, \theta) \int_0^\Delta ds = \mu(X_0, \theta) \Delta
$$

and

$$
\int_0^\Delta \sigma(X(s), \theta) \, dW(s) \approx \int_0^\Delta \sigma(X_0, \theta) \, dW(s) = \sigma(X_0, \theta) \int_0^\Delta dW(s) = \sigma(W(\Delta) - W(0)) = \sigma W(\Delta).
$$

Recalling from the property of an Itô integral (2.3) that $\int_0^\Delta dW(s)$ is normally distributed with mean zero and variance given by $\int_0^\Delta ds = \Delta$, the approximation to
\[ X(\Delta) = X_0 + \int_0^\Delta \mu(X(s), \theta) \, ds + \int_0^\Delta \sigma(X(s), \theta) \, dW(s) \]

is given by

\[ X(\Delta) \approx X_0 + \mu(X_0, \theta) \Delta + \sigma(X_0, \theta) \sqrt{\Delta} Z, \quad (3.1) \]

where \( Z \sim N(0, 1) \).

The approximated process in (3.1) leads to the approximate transition density

\[ p(X(\Delta)|X_0, \theta) \approx \xi(X(\Delta)|X_0, \theta), \]

defined to be

\[ \xi(X(\Delta)|X_0, \theta) = \phi(X(\Delta); X_0 + \mu(X_0, \theta) \Delta, \sigma^2(X_0, \theta) \Delta), \quad (3.2) \]

where \( \phi(x; m, \nu^2) \) is the density of a normal random variable with mean \( m \) and variance \( \nu^2 \) evaluated at \( x \). The error of the Euler-Maruyama method is \( O(\sqrt{\Delta}) \) [Shreve, 2004] in the sense of strong convergence. That is, for the true process \( X(t) \) and the discrete approximation of the process generated using the scheme in (3.1), say \( X^\Delta(t) \), we have

\[ E \left[ |X(t) - X^\Delta(t)| \right] \leq K \sqrt{\Delta}, \]

for \( 0 \leq t \leq T \) some \( K > 0 \).

**Higher-Order Approximations**

The Milstein approximation [Mil’shtein, 1975] adds a second-order term (via Itô’s lemma (2.12)) to make the approximation more accurate. The approximation is
obtained by first considering \( f(t, X(t)) = \mu(X(s), \theta) \) and \( f(t, X(t)) = \sigma(X(s), \theta) \)
which, via Itô’s Lemma, can be written as:

\[
d\mu(X(s), \theta) = \left[ \mu(X(t), \theta) \mu_x(X(t), \theta) + \frac{1}{2} \sigma^2(X(t), \theta) \mu_{xx}(X(t), \theta) \right] dt \\
+ \sigma(X(t), \theta) \mu_x(X(t), \theta) dW_t,
\]

and

\[
d\sigma(X(s), \theta) = \left[ \mu(X(t), \theta) \sigma_x(X(t), \theta) + \frac{1}{2} \sigma^2(X(t), \theta) \sigma_{xx}(X(t), \theta) \right] dt \\
+ \sigma(X(t), \theta) \sigma_x(X(t), \theta) dW_t,
\]

where

\[
\mu_x(X(t), \theta) = \frac{d}{dx} \mu(x, \theta) \bigg|_{x=X(t)},
\]
\[
\mu_{xx}(X(t), \theta) = \frac{d^2}{dx^2} \mu(x, \theta) \bigg|_{x=X(t)},
\]
\[
\sigma_x(X(t), \theta) = \frac{d}{dx} \sigma(x, \theta) \bigg|_{x=X(t)}, \quad \text{and}
\]
\[
\sigma_{xx}(X(t), \theta) = \frac{d^2}{dx^2} \sigma(x, \theta) \bigg|_{x=X(t)}.
\]

We can then plug the integral forms of these two quantities into

\[
X(\Delta) = X_0 + \int_0^\Delta \mu(X(s), \theta) \, ds + \int_0^\Delta \sigma(X(s), \theta) \, dW(s)
\]

to get

\[
X(\Delta) = X_0 + \int_0^\Delta \left\{ \mu(X_0, \theta) + \int_0^u \left[ \mu(X(t), \theta) \mu_x(X(u), \theta) \\
+ \frac{1}{2} \sigma^2(X(u), \theta) \mu_{xx}(X(u), \theta) \right] du \right\}
\]

24
\begin{align*}
+ \int_0^s \sigma(X(u), \theta) \mu_x(X(u), \theta) dW(u) \right\} ds \\
+ \int_0^\Delta \left\{ \sigma(X_0, \theta) + \int_0^s \left[ \mu(X(t), \theta) \right] \right. \\
+ \frac{1}{2} \sigma^2(X(u), \theta) \sigma_{xx}(X(u), \theta) \right] du \\
+ \int_0^s \sigma(X(u), \theta) \sigma_x(X(u), \theta) dW(u) \right\} dW(s).
\end{align*}

Note that from (2.9) and (2.10) many of these terms vanish, leaving us with the simpler expression given by

\begin{align*}
X(\Delta) &= X_0 + \int_0^\Delta \mu(X_0, \theta) ds + \int_0^\Delta \sigma(X_0, \theta) dW(s) \\
&\quad + \int_0^s \int_0^s \sigma(X(u), \theta) \sigma_x(X(u), \theta) dW(u) dW(s),
\end{align*}

where the first two integrals are the same as in the Euler-Maruyama scheme and we use the Euler approximation to the third integral, which allows us to write

\begin{align*}
\int_0^\Delta \int_0^s \sigma(X(u), \theta) \sigma_x(X(u), \theta) dW(u) dW(s) \\
&= \sigma(X_0, \theta) \sigma_x(X_0, \theta) \int_0^\Delta \int_0^s dW(u) dW(s) \\
&= \sigma(X_0, \theta) \sigma_x(X_0, \theta) \int_0^\Delta (W(s) - W_0) dW(s) \\
&= \sigma(X_0, \theta) \sigma_x(X_0, \theta) \left( \int_0^\Delta W(s) dW(s) - W_0 W_\Delta + W_0^2 \right)
\end{align*}
\[\sigma(X_0, \theta) \sigma_x(X_0, \theta) \int_0^\Delta W(s) \, dW(s) = \sigma(X_0, \theta) \sigma_x(X_0, \theta) \left( \frac{1}{2} W_\Delta^2 - \frac{\Delta}{2} \right).\]

The last step follows because

\[
\int_0^T W(t) \, dW(t) = \lim_{n \to 0} \sum_{i=1}^n \left[ \frac{1}{2} (W(t_i) + W(t_{i-1})) - \frac{1}{2} (W(t_i) - W(t_{i-1})) \right] \\
\times (W(t_i) - W(t_{i-1})) \\
= \lim_{n \to 0} \frac{1}{2} W(T)^2 - \frac{1}{2} \lim_{n \to 0} \sum_{i=1}^n (W(t_i) - W(t_{i-1}))^2 \\
= \frac{1}{2} W^2(T) - \frac{T}{2}. \tag{3.3}
\]

The difference of \(T/2\) between the Itô integral and the typical \(\int_0^T x \, dx = x^2/2\) result arises due to the fact that \(W(t)\) is not differentiable in \(t\) and the fact that the simple functions used to construct the Itô integral are evaluated at their left endpoints.

The Milstein approximation to the SDE is thus given by

\[X(\Delta) \approx X_0 + \mu(X_0, \theta) \Delta + \sigma(X_0, \theta) \sqrt{\Delta} Z + \frac{\Delta}{2} \sigma(X_0, \theta) \sigma_x(X_0, \theta) \left[ (Z^2 - 1) \right], \tag{3.4}\]

where \(Z \sim N(0, 1)\).

The approximation to the transition density based on the Milstein scheme, derived in Elerian [1998], is then given by

\[p(X(\Delta)|X_0, \theta) \approx \omega(X(\Delta)|X_0, \theta),\]
defined to be

\[ \omega(X(\Delta)|X_0, \theta) = \frac{e^{-\lambda/2}}{|A|\sqrt{2\pi}} V_\Delta^{-1/2} e^{-V \Delta/2} \cosh(\sqrt{V_\Delta}), \] (3.5)

where

\[ V_\Delta = \frac{X(\Delta) - B}{A}, \]

\[ \lambda = \frac{1}{\Delta \sigma_x^2(X_0, \theta)}, \]

\[ A = \frac{\sigma(X_0, \theta)\sigma_x(X_0, \theta)\Delta}{2}, \]

and

\[ B = -\frac{\sigma(X_0, \theta)}{2\sigma_x(X_0, \theta)} + X_0 + \mu(X_0, \theta)\Delta - \frac{\sigma(X_0, \theta)\sigma_x(X_0, \theta)\Delta}{2}. \]

The error of the Milstein method in terms of strong convergence is \( O(\Delta) \) [Shreve, 2004], smaller than that of the Euler method of \( O(\sqrt{\Delta}) \), since typically the sampling interval \( \Delta < 1 \).

The Ozaki method uses local linearization, which consists of approximating the drift of the process using a linear function rather than a constant as in the Euler-Maruyama scheme. Under the assumption that the diffusion term is constant, say \( \sigma \), the approximate process is generated by, conditionally on \( X_0 \), simulating \( X(\Delta) \) from a normal distribution with mean and variance given respectively by

\[ E_x = X_0 + \frac{\mu(X_0, \theta)}{\mu_x(X_0, \theta)} \left( e^{\mu_x(X_0, \theta)\Delta} - 1 \right) \] and \[ V_x = \sigma^2(X_0, \theta) e^{2K_x \Delta} - 1 \]

where

\[ K_x = \frac{1}{\Delta} \log \left( 1 + \frac{\mu(X_0, \theta)}{X(\Delta)\mu_x(X(\Delta), \theta)} \left( e^{\mu_x(X(\Delta), \theta)\Delta} - 1 \right) \right). \]

The approximation to the transition density based on the Ozaki scheme is then given by

\[ p(X(\Delta)|X_0, \theta) \approx \nu(X(\Delta)|X_0, \theta), \]
which is defined to be

\[ \nu (X(\Delta)|X_0, \theta) = \phi (X(\Delta); E_x, V_x). \quad (3.6) \]

Estimation of the transition density based on the above approximations has the advantage of being very general, but at the cost of being quite inaccurate for large time increments between observations. This is because the simple approximations to the Itô integrals become increasingly inaccurate as \( \Delta \), and therefore the distance between the limits of integration, grows [see, for example, Fuchs, 2013]. The maximum likelihood estimates based on the transition densities estimated using one of methods above are consistent as \( \Delta \rightarrow 0 \) (data are observed more frequently in time). However, in the more practical case of \( \Delta \) fixed and \( n \rightarrow \infty \) (more data are observed at the same sampling frequency), it can be shown that the MLEs are not consistent [Lo, 1988]. We now turn to the four groups of more recent approximation methods.

### 3.2 Solving the Forward Equation

The first of the four categories of more recent work done to approximate the transition density involves numerically solving the Kolmogorov forward (or Fokker-Planck) equation, which can be used to determine the stationary distribution of an SDE. In Lo [1988], the author demonstrates this method by first proving that the fact that the transition density \( p(X(\Delta)|X_0, \theta) \) satisfies the *Kolmogorov forward equation* given by

\[
\frac{dp}{dt}(X(\Delta)|X_0, \theta) = - \frac{d}{dx} \left[ \mu(X(\Delta), \theta)p(X(\Delta)|X_0, \theta) \right] \\
+ \frac{1}{2} \frac{d^2}{dx^2} \left[ \sigma^2(X(\Delta), \theta)p(X(\Delta)|X_0, \theta) \right]. \quad (3.7)
\]
is equivalent to

\[
\frac{dp}{dt}(X(\Delta)|X_0, \theta) = a(X(\Delta), \theta)p(X(\Delta)|X_0, \theta) + b(X(\Delta), \theta)\frac{dp}{dx}(X(\Delta)|X_0, \theta)
\]

\[+ c(X(\Delta), \theta)\frac{d^2}{dx^2}p(X(\Delta)|X_0, \theta), \tag{3.8}\]

where

\[
a(X(\Delta), \theta) = -\frac{d}{dx}\mu(X(\Delta), \theta) + \frac{1}{2} \frac{d^2}{dx^2}\sigma^2(X(\Delta), \theta),
\]
\[
b(X(\Delta), \theta) = -\mu(X(\Delta), \theta) + 2\sigma(X(\Delta), \theta)\frac{d}{dx}\sigma(X(\Delta), \theta), \text{ and}
\]
\[
c(X(\Delta), \theta) = \frac{1}{2}\sigma^2(X(\Delta), \theta).
\]

Equation (3.8) can then be numerically solved to provide an approximate transition density \(p(X(\Delta)|X_0, \theta)\). Poulsen [1999] provides a general procedure for solving this equation based on the Crank-Nicolson method, which is a numerical method to solve partial differential equations [Crank and Nicolson, 1947]. The earliest of the four classes of recent approximations, methods based on solving the forward equation have seen less development lately.

### 3.3 Closed-Form Expansions

Aït-Sahalia [2002a] provides a closed-form approximation to the likelihood (2.17), where the process is observed on regular intervals so that \(t_i = i\Delta, i = 0, \ldots, N\) (although details are provided for extending the method to irregularly sampled data).

The author begins by twice transforming the data. First the SDE \(\{X(t), t \geq 0\}\) is transformed into another SDE \(\{Y(t), t \geq 0\}\). The SDE is then further transformed from \(\{Y(t), t \geq 0\}\) into \(\{Z(t), t \geq 0\}\). Notationally, we denote the transition densities of the transformed data as \(p_Y(\cdot|\cdot, \cdot)\) and \(p_Z(\cdot|\cdot, \cdot)\), respectively. The transformation
from $X$ into $Y$ so that the process has unit diffusion is the Lamperti transform, given by (2.13).

Next, the author defines the pseudo-normalized increment of $Y$ as

$$Z(t) \equiv \Delta^{-1/2}(Y(t) - y_0). \quad (3.9)$$

While further details may be found in the paper, the author heuristically motivates this transformation by describing $\Delta$ as playing the role of the sample size $N$ in the central limit theorem (CLT). Properly standardizing the data will allow us to determine the limiting distribution as $\Delta \to 0$. Similar to practical applications of the CLT when $N$ is not infinity, corrections to the $\phi(x; 0, 1)$ approximation can be made by considering higher order terms, say $J$ of them for $J > 0$. The author motivates the choice of higher order terms based on Hermite polynomials due to the orthogonality of these polynomials with respect to the $\phi(x; 0, 1)$ term.

Define the order $J$ Hermite expansion of $p_Z(z\Delta|y_0; \theta)$ as

$$p_Z^{(J)}(z\Delta|y_0; \theta) \equiv \phi(z) \sum_{j=0}^{J} \eta_Z^{(j)}(\Delta, y_0; \theta)H_j(z), \quad (3.10)$$

where $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$, $H_j(z)$ is the probabilists’ Hermite polynomial given by

$$H_j(z) \equiv e^{z^2/2} \frac{d^j}{dz^j} \left[ e^{-z^2/2} \right], \quad \text{for } j \geq 0. \quad (3.11)$$

In the Hermite expansion given above, the zeroth order coefficient $\eta_Z^{(0)}(\Delta, y_0, \theta) = 1$ and the $j^{th}$ order coefficient (for $j \geq 1$) in the Hermite expansion is

$$\eta_Z^{(j)}(\Delta, y_0, \theta) = \frac{1}{j!} \int_{-\infty}^{\infty} H_j(z)p_Z(Z(\Delta)|y_0; \theta)dz$$

$$= \frac{1}{j!} \int_{-\infty}^{\infty} H_j(\Delta^{-1/2}(Y - y_0)) p_Y(Y(\Delta)|y_0, \theta)dy$$

30
These coefficients, \( \eta^{(j)}_Z \), can be estimated using Monte Carlo integration or, as the author illustrates, using a Taylor series expansion in \( \Delta \):

\[
E[f(Y(\Delta), y_0)|Y(0) = y_0] = \sum_{k=0}^{K} A^k(\theta) f(Y(\Delta), y_0) \frac{\Delta^k}{k!} + \text{remainder},
\]

where \( A(\theta) \) is the infinitesimal generator of the diffusion \( Y \), defined by

\[
f \mapsto A(\theta) f \equiv \mu_Y(\cdot; \theta) \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2}.
\]

Equations (4.4)–(4.9) in A¨ıt-Sahalia [2002a] provide the first Taylor series up to order \( K = 3 \) for the first seven Hermite coefficients \( (j = 0, \ldots, 6) \) as \( \eta^{(0)}_Z = 1 \) and

\[
\eta^{(1,3)}_Z = -\mu_Y \Delta^{1/2} - \frac{(2\mu_Y \mu_Y^{[1]} + \mu_Y^{[2]}) \Delta^{3/2}}{4} - \left( \frac{4\mu_Y^{[1]} \mu_Y^{[2]} + 4\mu_Y^{[1]} \mu_Y^{[2]} + 6\mu_Y^{[1]} \mu_Y^{[2]} + 4\mu_Y \mu_Y^{[3]} + \mu_Y^{[4]}}{24} \right) \Delta^{5/2},
\]

\[
\eta^{(2,3)}_Z = \frac{ \left( \mu_Y^2 + \mu_Y^{[1]} \right) \Delta}{2} + \frac{ \left( 6\mu_Y^{[1]} \mu_Y^{[2]} + 4\mu_Y \mu_Y^{[3]} + 7\mu_Y \mu_Y^{[2]} + 2\mu_Y^{[3]} \right) \Delta^2}{12},
\]

\[
+ \left( \frac{28\mu_Y^2 \mu_Y^{[1]} + 28\mu_Y^2 \mu_Y^{[2]} + 16\mu_Y^{[1]} + 16\mu_Y^3 \mu_Y^{[2]} + 88\mu_Y \mu_Y^{[1]} \mu_Y^{[2]} + 21\mu_Y^{[2]} + 32\mu_Y^{[1]} \mu_Y^{[3]} + 16\mu_Y \mu_Y^{[4]} + 3\mu_Y^{[5]} \right) \Delta^{3/2} / 96,
\]

\[
\eta^{(3,3)}_Z = -\frac{ \left( \mu_Y^3 + 3\mu_Y \mu_Y^{[1]} + \mu_Y^{[2]} \right) \Delta^{3/2}}{6} - \left( \frac{12\mu_Y^3 \mu_Y^{[1]} + 28\mu_Y \mu_Y^{[1]} + 22\mu_Y^2 \mu_Y^{[2]} + 24\mu_Y^{[1]} \mu_Y^{[2]} + 14\mu_Y \mu_Y^{[3]} + 3\mu_Y^{[4]} \right) \Delta^{5/2},
\]

\[
\eta^{(4,3)}_Z = \frac{ \left( \mu_Y^4 + 6\mu_Y^2 \mu_Y^{[1]} + 3\mu_Y^{[1]} + 4\mu_Y \mu_Y^{[2]} + \mu_Y^{[3]} \right) \Delta^2}{24} + \left( \frac{20\mu_Y^4 \mu_Y^{[1]} + 50\mu_Y^3 \mu_Y^{[2]} + 100\mu_Y^2 \mu_Y^{[1]} + 50\mu_Y \mu_Y^{[3]} + 23\mu_Y \mu_Y^{[4]} + 180\mu_Y \mu_Y^{[1]} \mu_Y^{[2]} + 40\mu_Y^{[1]} + 34\mu_Y^{[2]} + 52\mu_Y^{[1]} \mu_Y^{[3]} + 4\mu_Y^{[5]} \right) \Delta^{3/2} / 240,
\]
\[ \eta_{Z}^{(5,3)} = -\left( \mu_{Y}^{5} + 10\mu_{Y}^{3}\mu_{Y}^{[1]} + 15\mu_{Y}\mu_{Y}^{[1]} + 10\mu_{Y}^{2}\mu_{Y}^{[2]} \\
+ 10\mu_{Y}[\mu_{Y}^{[2]} + 5\mu_{Y}\mu_{Y}^{[3]} + \mu_{Y}^{[4]}]) \Delta^{5/2}/120, \right. \\
\eta_{Z}^{(6,3)} = \left( \mu_{Y}^{6} + 15\mu_{Y}^{4}\mu_{Y}^{[1]} + 15\mu_{Y}^{3}\mu_{Y}^{[2]} + 20\mu_{Y}\mu_{Y}^{[1]} + 15\mu_{Y}\mu_{Y}^{[3]} + 45\mu_{Y}^{2}\mu_{Y}^{[1]} \\
+ 10\mu_{Y}^{[2]} + 15\mu_{Y}^{2}\mu_{Y}^{[3]} + 60\mu_{Y}\mu_{Y}^{[1]} + 6\mu_{Y}\mu_{Y}^{[4]} + \mu_{Y}^{[5]} \right) \Delta^{3}/720, \]

where the \(i^{th}\) order derivative of the drift term is given by

\[ \mu_{Y}^{[i]} = \frac{\partial^{i} \mu(y_{0}; \theta)}{\partial y_{0}^{i}}. \]

The Taylor series approximation for the Hermite coefficients can then be used to approximate the Hermite approximation to the transition density as

\[ p_{Z}^{(J)}(z_{\Delta}|y_{0}, \theta) \approx p_{Z}^{(J,K)}(z_{\Delta}|y_{0}, \theta), \]

which is defined to be

\[ p_{Z}^{(J,K)}(z_{\Delta}|y_{0}, \theta) = \phi(z) \left[ 1 + \sum_{j=1}^{J} \eta_{Z}^{(j,K)}(\Delta, y_{0}; \theta)H_{j}(z) \right]. \]

In contrast to the CLT, where convergence is understood to mean that the series with a fixed number of corrective terms \(J\) converges as the number of data points \(N \rightarrow \infty\), here we are interested in what happens as \(J \rightarrow \infty\). However, the series will actually diverge as more corrective terms are added, unless the density is “close enough” to a standard normal to begin with. Formally, “close enough” means that for a density \(p(v)\) to be expanded around a \(N(0,1)\), it must have thin enough tails for

\[ \int \exp(v^{2}/2)p'(v)^{2}dv < \infty, \]

where \(p'(v)\) is the derivative of the density with respect to \(v\). (Further detail on this condition may be found in Ait-Sahalia [2002a].) In general, \(p(X(\Delta)|X_{0}, \theta)\) cannot be
approximated for fixed $\Delta$ and the author states that the class of densities for which the “Hermite expansions will converge in the sense of adding more terms ($J$ increases with $\Delta$ fixed) is quite limited”. However, the specific transformed variable $Z$ given by (3.9) has a density $p_Z(\cdot,\cdot)$ that belongs to this class.

In Aït-Sahalia [2008] the author extends the closed-form approximation to the case of those multivariate diffusions which are reducible; that is, can be transformed to have diffusion coefficient equal to one.

3.4 Sequential Monte Carlo

Sequential Monte Carlo (SMC), sometimes referred to as simulated maximum likelihood estimation (SMLE), is the third group of methods used to approximate the transition density. In contrast to many of the other procedures, it does not require transforming (2.5) into an SDE of unit diffusion and it can be made arbitrarily accurate at the expense of more computation. As discussed in Chapter 2, for large values of $\Delta$ (which we can expect to have in practice) the Euler approximation is inaccurate. SMC aims to remedy this by partitioning the interval $[0,\Delta)$ into $K$ subintervals with endpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_K = \Delta$, such that $\tau_j - \tau_{j-1} = \Delta/K$, $j = 1, 2, \ldots, K$. Then, consider the SDE at the unobserved points $X(\tau) = (X(\tau_1),X(\tau_2),\ldots,X(\tau_{K-1}))$. The SMC approximation is based on the following representation of the transition density:

$$p(X(\Delta)|X_0,\theta) = \int p(X(\Delta),X(\tau)|X_0,\theta) \lambda(dX(\tau))$$

$$= \int \prod_{k=1}^K p(X(\tau_k)|X(\tau_{k-1}),\theta) \lambda(dX(\tau)),$$

where $\lambda$ denotes the Lebesgue measure on $\mathbb{R}$. 

33
Replacing the transition densities (which are over a shorter distance in time than the original transition density) in (3.14) by their Euler approximations yields the discretized transition density [see Kloeden and Platen, 1992] defined as

\[
p^{(K)}(X(\Delta)|X_0, \theta) = \int \prod_{k=1}^{K} \xi(X(\tau_k)|X(\tau_{k-1}), \theta) \lambda(dX(\tau)).
\] (3.15)

In Pedersen [1995b], the author provides sufficient conditions for the existence of \( p^{(K)}(X(\Delta)|X_0, \theta) \) over the support of \( X \) for \( K \geq 1 \) and the regularity conditions that ensure

\[
\lim_{K \to \infty} p^{(K)}(X(\Delta)|X_0, \theta) = p(X(\Delta)|X_0, \theta),
\] (3.16)

where the above limit is viewed in the \( L^1(\lambda) \) sense.

The author also shows that the discretized log-likelihood of the data based on (3.15), given by

\[
l^{(K)}(\theta|X) = \sum_{i=1}^{N} p^{(K)}(X_i|X_{i-1}, \theta),
\]

converges to the true log-likelihood as an immediate consequence of (3.16). That is,

\[
\lim_{K \to \infty} l^{(K)}(\theta|X) = l(\theta|X) \text{ in } L^1(\lambda),
\]

where \( l(\theta|X) \) is shorthand for the logarithm of \( L(\theta|X) \) defined in (2.17).

In Bally and Talay [1995], the bias of \( p^{(K)}(X(\Delta)|X_0, \theta) \) is shown to be of order \( O(1/K) \). So if the integral in (3.15) were available in closed-form we could choose \( K \) large enough to reduce this discretization error to a sufficiently small level and approximate the log-likelihood accordingly. However, as the integral is not available in closed-form, we resort to Monte Carlo methods to estimate it.

Early work by Ogawa [1995], Hurn and Lindsay [1999], and Nicolau [2002] provides a straightforward method using simulated values from the Euler approximation.
to obtain a kernel density estimate of the transition density. As noted in Brandt and Santa-Clara [2002b], "this approach suffers from the usual problems with nonparametric density estimation: a slower convergence rate (in the number of simulations) and the curse of dimensionality." That is, as the number of variables increases, the convergence rate of these estimators to their asymptotic distributions deteriorates exponentially.

We can use importance sampling to estimate the integral in (3.15), rewriting it as

\[
p^{(K)}(X(\Delta)|X_0, \theta) = \int \prod_{k=1}^{K} \xi(X(\tau_k)|X(\tau_{k-1}), \theta) \frac{q(X(\tau_1), \ldots, X(\tau_{k-1}))}{q(X(\tau_1), \ldots, X(\tau_k))} \times q(X(\tau_1), \ldots, X(\tau_{k-1})) d\lambda(X(\tau_1), \ldots, X(\tau_{k-1}))
\]

\[
= E_q \left[ \prod_{k=1}^{K} \xi(X(\tau_k)|X(\tau_{k-1}), \theta) \right]
\]

\[
= E_q [R^{(K)}],
\]

where \(E_q\) refers to the expectation with respect to the importance sampler \(q(\cdot)\), and \(R^{(K)}\) is defined as:

\[
R^{(K)} = \prod_{k=1}^{K} \frac{\xi(X(\tau_k)|X(\tau_{k-1}), \theta)}{q(X(\tau))}.
\]

We use a classical Monte Carlo estimator for this expectation, given by

\[
p^{(K,M)}(X(\Delta)|X_0, \theta) = \frac{1}{M} \sum_{m=1}^{M} R^{(K)}_m,
\]

where the variates \(R^{(K)}_m\) \((m = 1, \ldots, M)\) are calculated based on \(M\) independent and identically distributed (IID) draws \(X_m(\tau) = (X_m(\tau_1), \ldots, X_m(\tau_{K-1})) \sim q(\cdot)\). The
construction of the estimated log-likelihood is a straightforward application of the Markov property similar to that given in (2.17) as
\[
l^{(K,M)}(\theta|X) \equiv \sum_{i=1}^{N} \log p^{(K,M)}(X_i|X_{i-1}, \theta).
\]

As noted in Stramer and Yan [2007b] the Monte Carlo error, defined to be
\[
p^{(K,M)}(X(\Delta)|X_0, \theta) - p^{(K)}(X(\Delta)|X_0, \theta)
\]
is of order \(O(\sqrt{\text{Var}_q(R^{(K)})/M})\) assuming that \(\text{Var}_q(R^{(K)})\) exists.

It is clear from (3.17) and (3.18) that selecting a good importance sampling density \(q(\cdot)\) is critical. The optimal choice of \(q(\cdot)\) is the true joint density of \(X(\tau)\) given \(X_0\) and \(X(\Delta)\) [Stramer and Yan, 2007a], which is unavailable, as it depends on the transition density that we are trying to estimate. We now discuss various choices of \(q(\cdot)\).

### 3.4.1 Pedersen’s Method

In Pedersen [1995a] and independently in an earlier version of Brandt and Santa-Clara [2002b], the authors choose to simulate \(X(\tau)\) sequentially using the Euler approximation. That is, they specify
\[
q(X(\tau)) = \prod_{k=1}^{K-1} \xi(X(\tau_k)|X(\tau_{k-1}), \theta).
\]

In that case, we have by cancellation of terms,
\[
R^{(K)} = \frac{\prod_{k=1}^{K} \xi(X(\tau_k)|X(\tau_{k-1}), \theta)}{q(X(\tau))} = \frac{\prod_{k=1}^{K} \xi(X(\tau_k)|X(\tau_{k-1}), \theta)}{\prod_{k=1}^{K-1} \xi(X(\tau_k)|X(\tau_{k-1}), \theta)} = \xi(X(\Delta)|X(\tau_{K-1}), \theta).
\]
Then (3.18) can be rewritten as

\[ p^{(K,M)}(X(\Delta)|X_0, \theta) = \frac{1}{M} \sum_{m=1}^{M} \xi (X(\Delta)|X_m(\tau_{K-1}), \theta) , \] (3.20)

where \( X_m(\tau_{K-1}) \) is obtained by sequentially simulating

\[ X_m(\tau_1)|X_0, X_m(\tau_2)|X_m(\tau_1), \ldots, X_m(\tau_{K-1})|X_m(\tau_{K-2}) , \]

each using the Euler scheme given by (3.2) for \( m = 1, \ldots, M \).

Although appealing due to its simplicity, this sampler has been criticized for its inefficiency. As noted in Stramer and Yan [2007b], \( R^{(K)} \) in this method may have unbounded variance, causing the sampler to be inefficient.

### 3.4.2 Durham’s and Gallant’s Method

Durham and Gallant [2002] propose two more efficient Monte Carlo samplers, which benefit from simulating the \( X(\tau) \) values conditionally on both the initial and terminal points, instead of only the initial point as in the previous method. The Brownian bridge sampler is defined to be the Euler approximation of the solution \( H_t \) to

\[ dH_t = \frac{X(\Delta) - H_t}{\Delta - t} dt + \sigma(H_t, \theta)dW_t, \quad 0 \leq t \leq \Delta, \quad H_0 = X_0. \] (3.21)

For a constant \( \sigma(\cdot) \), \( H_t \) can be shown to be a Brownian bridge (2.1) on \([0, \Delta]\) from \( X_0 \) to \( X(\Delta) \) [Stramer and Yan, 2007a].

The modified Brownian bridge sampler is introduced based on the recursion

\[ X(\tau_k) = X(\tau_{k-1}) + \frac{X(\tau_K) - X(\tau_{k-1})}{K - k + 1} \]
\[ + \sqrt{\frac{K - k}{K - k + 1}} \sqrt{\frac{\Delta}{K}} \sigma(X(\tau_{k-1}), \theta) Z_k, \quad \text{for } k = 1, \ldots, K - 1 \] (3.22)
where $X(\tau_0) = X_0$, $X(\tau_K) = X(\Delta)$, and $Z_k^{HD} \sim N(0, 1)$. Note that this sampler is identical to (3.21) except for the $(K - k)/(K - k + 1)$ term in the variance. The authors state “it is not entirely obvious that this should be the case, [but] we will see that this modification results in much better performance”. To lend support to the heuristic arguments justifying (3.22), Stramer and Yan [2007a] show that when $\sigma(\cdot)$ is constant, the modified Brownian bridge is also exactly a Brownian bridge on $[0, \Delta]$ from $X_0$ to $X(\Delta)$.

As discussed in Chib and Shephard [2002], $R^{(K)}$ based on this scheme can be interpreted as a simple Euler approximation multiplied by the expected value of the ratio of two predictive densities. By simulating conditional on $X(\Delta)$ in addition to $X_0$, enormous gains in efficiency are achieved. Furthermore, guidance is available for the somewhat arbitrary choices of $K$ and $M$. We thus choose to use the modified Brownian Bridge sampler for estimation of the log-likelihood at a fixed $\theta$ and we set $M = K^2$, as Stramer and Yan [2007b] show that this choice is computationally optimal for a “fixed large amount of computer time”.

### 3.4.3 Other SMC Methods

Elerian et al. [2001] criticized Pedersen’s method (3.20) for its inefficiency and proposed a computationally intensive method of sampling $X(\tau)$ from a multivariate Normal or $t$ distribution based on a second-order Taylor expansion. In Golightly and Wilkinson [2006] and Golightly and Wilkinson [2008], the authors generalize the Durham and Gallant sampler to fully observed and partially observed multivariate diffusion models, respectively. Lindström [2012] introduces a sampler which is a combination of the Pedersen sampler and the Durham and Gallant sampler in an
attempt to remedy the shortcomings of each. These shortcomings were outlined in Fearnhead [2008]. Lin et al. [2010] introduces a sampling scheme guided by resampling based on importance weights determined using backwards pilot simulations from the end observation.

Although originally developed outside of the SMC context, the exact method that is described in the following section can be thought of as a particular choice of $q(\cdot)$.

3.5 Exact Algorithm

Exact simulation [Beskos and Roberts, 2005, Beskos et al., 2006a, 2008] presents an opportunity to sample from a $q(\cdot)$ that is exactly $p(\mathbf{X}(\tau)|\mathbf{X}_0, \theta)$, but adds another layer of computational complexity [Bladt and Sørensen, 2014]. Similar to Pedersen’s method (3.20), the exact simulation method ignores $X(\Delta)$ in drawing variates, and thus suffers from efficiency issues. We begin with an explanation of the first version of the exact algorithm (EA1), briefly discuss the extensions (EA2 and EA3), and then discuss estimating the transition density based on exact simulation.

In Beskos and Roberts [2005], or “EA1”, the authors develop the initial method to generate exact skeletons of the path between observed points. The paper begins with a slightly different version of the usual rejection sampler used to sample from some difficult density $f(x)$. Let $g(x)$ be another density which is easy to calculate and sample from so that

$$\frac{f(x)}{g(x)} \leq \mu,$$

where $\mu > 1$. The usual rejection sampler proceeds by

1. Sample $Y \sim g(x)$.  

39
2. Sample $U \sim \text{Unif}(0,1)$.

3. If $U > \frac{f(Y)}{\mu g(Y)}$, return to step 1.

4. Otherwise, set $X = Y$ and consider $X$ to be an element of the sample. Return to step one until the sample is of desired size.

The authors use the following scheme to sample from some probability measure $\mu$ that is absolutely continuous with respect another probability measure $\nu$ and $f \equiv \epsilon \frac{d\mu}{d\nu} \leq 1$, $\nu$-almost surely for some $\epsilon > 0$. The scheme does not imply the ordering inherent in the usual rejection sampler (by first sampling $Y$ and then sampling $U$) and is “more easily understood when imagining that the proposed variate actually succeeds the decision variate” [Beskos and Roberts, 2005].

**Rejection Sampling**

Let $(Y_n, I_n)_{n \geq 1}$ be a sequence of iid random elements taking values in $S \times \{0, 1\}$ such that $Y_1 \sim \nu$ and $P[I_1 = 1 | Y_1 = y] = f(y)$ for all $y \in S$.

Define $\tau = \min\{i \geq 1 : I_i = 1\}$. Then $P[Y_\tau \in dy] = \mu(dy)$.

The process $X$ is assumed to have unit diffusion (otherwise we can use (2.13) to produce a process with unit diffusion). Let $\mathcal{Q}$ be the probability measure $X$ induces on continuous functions and the associated $\sigma$-field and let $\mathcal{W}$ be the Wiener measure on the same space. Then Girsanov’s formula (2.15) yields:

$$G(X) \equiv \frac{d\mathcal{Q}}{d\mathcal{W}}(X) = \exp \left( \int_0^T \mu(X_s) dX_s - \frac{1}{2} \int_0^T \mu^2(X_s) ds \right).$$  \hspace{1cm} (3.23)
The first condition that the authors assume is that $\mu(\cdot)$ is everywhere differentiable, which allows them to write:

$$G(X) = \exp \left( A(X(t)) - A(X_0) - \frac{1}{2} \int_0^T (\mu^2(X_s) + \mu'(X_s))ds \right),$$

where

$$A(u) = \int_0^u \mu(y)dy.$$  

The authors note that performing rejection sampling using Brownian candidates would likely require $A(\cdot)$ to be bounded. To get around this, they introduce the biased Brownian motion $\overline{BM}$, defined to be a Brownian Motion $\{W_t; 0 \leq t \leq T\}$ with the condition that $W_T = \rho$, where $\rho \sim h : \mathbb{R} \mapsto [0, \infty)$. Then, denoting the probability measure induced by $\overline{BM}$ as $Z$, it can be shown that

$$\frac{dZ}{d\overline{W}}(X) = \frac{h(X(t))}{(1/\sqrt{2\pi T}) \exp (X(t)^2/(2T))},$$

so that

$$\frac{dQ}{dZ}(X) \propto \exp \left( A(X(t)) - X(t)^2/2T - \frac{1}{2} \int_0^T (\mu^2(X_s) + \mu'(X_s))ds \right) / h(X(t)).$$

The second condition assumed in this paper is that

$$c \equiv \int_{\mathbb{R}} \exp \left( A(u) - u^2/2T \right) du < \infty.$$  

This assumption allows them to choose

$$h(u) = \exp \left( A(u) - u^2/2T \right) / c,$$

so that

$$\frac{dQ}{dZ}(X) \propto \exp \left( -\frac{1}{2} \int_0^T (\mu^2(X_s) + \mu'(X_s))ds \right).$$
The third, and most problematic, assumption is that the function in the exponent is bounded. In other words, there exist $k_1, k_2 \in \mathbb{R}$ such that

$$k_1 \leq \frac{1}{2} (\mu(u) + \mu'(u)) \leq k_2.$$ 

Defining

$$\psi(u) \equiv \frac{1}{2} (\mu^2(u) + \mu'(u)) - k_1,$$

the authors write:

$$\frac{dQ}{dZ}(X) \propto \exp \left( - \int_0^T \psi(X_s) ds \right). \quad (3.28)$$

The assumptions above allow the authors to define the ideal rejection sampler given by

1. Sample a complete, continuous path of $BM, \omega \sim Z$.
2. Compute $\int_0^T \psi(\omega_s) ds$.
3. Sample $I$ from a Bernoulli with $P[I = 1|\omega] = \exp \left( - \int_0^T \psi(\omega_s) ds \right)$.
4. If $I = 0$, return to 1.
5. Output $\omega$.

However, this sampler will be impossible due to the fact that the entire path $\omega$ cannot be simulated in practice. The exact algorithm allows the first two steps to be circumvented, after which steps 3 and 4 are carried out exactly on a finite skeleton of $\omega$. A skeleton is a set of observations (in discrete time) from $\omega$ which are exact in the sense that they are free from discretization error. The practical version of the rejection sampler is given by
Exact Algorithm 1

1. Initiate a path of the biased Brownian motion, $BM$. Set $\omega(0) = 0$ and draw $\omega(T) \sim h$.

2. Draw $U \sim \text{Unif}(0,1)$. Set $i = 0$.

3. Draw $(V, W) \sim \text{Unif}[(0, T) \times (0, \frac{1}{T})]$. Set $i = i + 1$.

4. Construct $\omega(V)$ given the currently simulated instances of $\omega$.

5. If $\psi(\omega(V)) < W$ or $U > \frac{1}{i}$ then

   If $i$ is even, set $I = 0$ and go to 1.

   If $i$ is odd, set $I = 1$ and go to 6.

5. Otherwise, go to 3.

6. Output the currently simulated instances of $\omega$.

To give some intuition behind step 3, the authors note that for $0 \leq \psi(u) \leq T^{-1}$, an event of probability $\int_0^T \psi(u)du$ can be constructed from drawing $(V, W) \sim \text{Unif}[(0, T) \times (0, T^{-1})]$ and considering the event $\{\psi(V) \geq W\}$. ($T$ must be chosen so that this is true, but multiple skeletons may be merged together if the desired $T > 1/(k_2 - k_1)$.) Step 5 is motivated by the fact that the idea in step 3 can be extended to

$$\exp\left(-\int_0^T \psi(u)du\right),$$
by a Taylor series expansion that depends upon a countable sequence of events of the probability in the display above. In addition, the event can be expressed as a union of a sequence of increasing events and an intersection of another sequence of decreasing events.

Depending on the context, a possible computational disadvantage of this algorithm is the restriction that \( T \leq 1/(k_2 - k_1) \). In other words, the length of the “exact” path that can be simulated is bounded. For the sine example that follows, this bound was 8/9. While multiple skeletons can be merged to get a path of the desired length, the requirement to know the ending point of the previous skeleton before generating the next one may cause computational problems for very large \( T \). In the example presented by the authors, they merged skeletons on \([0,8/9]\) and \([8/9,1]\) to compare their method with the Euler method and found that their method was “superior even in terms of computational time.”

The second, and biggest, disadvantage of this model is the restrictiveness of the assumptions on \( X(t) \). While the first two conditions are fairly common in the literature, the most problematic condition is the third one, that

\[
0.5 \left( \mu^2(X(t), \theta) + \mu'(X(t), \theta) \right)
\]

must be bounded below and above for all \( X(t) \in \mathbb{R} \). Even the commonly used Ornstein-Uhlenbeck model does not meet this requirement. To illustrate this approach, consider the example model

\[
dX(t) = \sin(X(t))dt + dW(t). \tag{3.29}
\]

For this model,

\[
\frac{1}{2} \left( \mu^2(u) + \mu'(u) \right) = \frac{1}{2} \left( \sin^2(u) + \cos(u) \right) \in \left[-\frac{1}{2}, \frac{5}{8}\right], \quad \forall u \in \mathbb{R},
\]
so the conditions of Assumption 3 are met for $k_1 = -1/2$ and $k_2 = 5/8$.

To investigate the behavior of (3.29), we attempt to use the following result [see, for example, Herbei and Berliner, 2014] to determine the stationary distribution of (3.29). Denoting the stationary distribution as $p(x)$ the Fokker-Planck equation gives

$$0.5 \frac{d}{dx} (\sigma^2(x, \theta)p(x)) = \mu(x, \theta)p(x),$$

which yields the general solution (with $c$ being the necessary integration constant):

$$p(x) = \frac{1}{c\sigma^2(x, \theta)} \exp \left( \int_0^x \frac{2\mu(z, \theta)}{\sigma^2(z, \theta)} dz \right).$$

Then attempting to use (3.31) to find the stationary distribution for $X(t)$ yields

$$p(x) = \frac{1}{c} \exp \left( \int_0^x 2\sin(z) dz \right) = \frac{1}{c} \exp \{2 [1 - \cos(x)] \}.$$

But attempting to solve for

$$c = \int_{-\infty}^{\infty} \exp \{2 [1 - \cos(x)] \} dx,$$

we find that the integral does not exist:

$$\infty = \int_{-\infty}^{\infty} \exp(-2) \leq \int_{-\infty}^{\infty} \exp \{2 [1 - \cos(x)] \} dx \leq \int_{-\infty}^{\infty} \exp(2) = \infty.$$

So, for this model there is no stationary distribution.

If $X(t)$ follows (3.29), the process will get trapped between multiples of $2\pi$ for extended periods of time. To illustrate, consider the interval $(0, 2\pi)$. Since $\sin(X(t)) > 0$ for $X(t) \in (0, \pi)$ and $\sin(X(t)) < 0$ for $X(t) \in (\pi, 2\pi)$, the process will have a tendency to revert to $\pi$. However, due to the randomness contributed by the Brownian Motion, occasionally $X(t)$ will enter $(-2\pi, 0)$ or $(2\pi, 4\pi)$, at which point the drift term will push the process toward $-\pi$ or $3\pi$, respectively.
Figure 3.1: Euler approximation (with $\delta = 0.01$) of $dX_t = \sin(X_t)dt + dW_t$ process observed until time $T = 100$ (top) and $T = 10000$ (bottom). These figures illustrate the behavior of the process used in Beskos and Roberts [2005] to demonstrate Exact Algorithm 1 (EA1).

The odd behavior of this model is illustrated graphically via simulation for observations spaced 0.01 in time, with $T = 100$ and $T = 10000$ in Figure 3.1.

In Beskos et al. [2006a], or “EA2”, the authors mainly attempt to relax the problematic condition from EA1 and improve efficiency. In EA1, $\mu^2(u) + \mu'(u)$ needed to be bounded from above and from below. However, in EA2, while $\mu^2(u) + \mu'(u)$ still
needs to be bounded from below, it is only required that either

\[ \limsup_{u \to \infty} \mu^2(u) + \mu'(u) < \infty, \text{ or} \]
\[ \limsup_{u \to -\infty} \mu^2(u) + \mu'(u) < \infty. \]

Beskos et al. [2006a] also introduces the Poisson process used to determine the number of points to be generated, which provides a simpler EA1, denoted by EA1a.

The advantages to EA2 are that when the restrictive conditions of EA1 are met, it provides a more efficient rejection sampling scheme using a Poisson process. Also, by relaxing the third condition from EA1, EA2 allows for the consideration of a wider class of models. One disadvantage of EA2 has to do with computational time. Now that the exact algorithm depends on the minimum of the path, which can be arbitrarily small for a given \( \omega \), the authors state that it is impossible to uniformly bound \( E[D|\omega] \) over \( \omega \), where \( D \) is the number of Poisson process points needed to make a decision about whether to accept or reject a proposed path. The authors suggest preliminary runs to estimate how long the algorithm will take to run, and if necessary, redesigning the algorithm. They note that in the example above, they needed to adjust \( T \) according to the values of the parameters.

In Beskos et al. [2008], or “EA3”, the authors essentially remove the assumption from EA2 that either \( \limsup_{u \to \infty} (\mu^2(u) + \mu'(u)) \) or \( \limsup_{u \to -\infty} (\mu^2(u) + \mu'(u)) \) is not \( +\infty \). They do this by considering a finite-dimensional RV \( \Upsilon \) and positive function \( r(\cdot) \) such that

\[ \sup_{s \in [0,t]} \psi(X_s) \leq r(\Upsilon) < \infty. \tag{3.32} \]
Under the conditions of EA1, one can choose

\[ r(u) \equiv r = \sup_{z \in \mathbb{R}} \psi(z). \]

Under the conditions of EA2,

\[ \Upsilon = m = \inf \{ X_s; s \in [0, t] \} \] and

\[ r(u) = \sup_{z \in [u, \infty)} \psi(z), u \in \mathbb{R} \]

will suffice.

The main advantage of EA3 is that it significantly widens the class of diffusions that can be exactly simulated. The construction also allows for straightforward simulation of hitting times by properly specifying the layers. A disadvantage of EA3 is the increased complexity of the algorithm compared to the previous two versions. This increased complexity is due to the rejection sampler necessary to get the layered Brownian bridge. As stated in Sermaidis et al. [2013], “EA1 and EA2 should still be preferred whenever possible; simulating a layered Brownian bridge is a non-trivial task”. In that article, they state that EA3 is roughly 10 times slower than EA1.

3.5.1 Using EA to Estimate Transition Densities

In Beskos et al. [2006b], the authors demonstrate how EA can be used for Monte Carlo estimation of the likelihood by conditioning on the skeletons. That is, use the fact that, for a skeleton \( S \), the transition density can be rewritten as

\[ p(X_\Delta|X_0; \theta) = E_S [p(X_\Delta|X_0; \theta, S)], \quad (3.33) \]

where the expectation is taken with respect to \( S \) and \( p(X_\Delta|X_0; \theta, S) \) is the usual transition density conditioned on \( S \).
As summarized in Fuchs [2013], the estimation procedure (using EA1) proceeds as follows.

1. For \( m = 1, \ldots, M \) do:
   
   (a) Use EA to draw a skeleton \( S = ((0, X_0), (t_1, X^*_1), \ldots (t_k, X^*_k), (T, X_T)) \) on \([0, T]\), where \( T > \Delta \).

   (b) Identify
   
   \[
   t_l = \max \{ t_i | t_i < t, i = 0, \ldots, k + 1 \} 
   \]
   
   and \( t_r = \min \{ t_i | t_i > t, i = 0, \ldots, k + 1 \} \),
   
   where \( t_0 = 0 \) and \( t_{k+1} = T \) and the corresponding \( X_l \) and \( X_r \) are chosen to be the values from the skeleton corresponding to the times \( t_l \) and \( t_r \).

   (c) Compute
   
   \[
   p^m = p(X_\Delta | X_0; \theta, X_l, X_r) = \phi \left( X_\Delta; X_l + \frac{t - t_l}{t_r - t_l} (X_r - X_l), \frac{(t - t_l)(t_r - t)}{t_r - t_l} \right).
   \]

2. Estimate \( p(X_\Delta | X_0; \theta) \) by \( \sum_{m=1}^{M} p^m \).

The calculation of \( p(X_\Delta | X_0; \theta, X_l, X_r) \) follows from the fact that conditioned on the skeleton, the diffusion has the same law as a Brownian bridge, as shown in Beskos and Roberts [2005]. Note that this method will often be computationally complex, as simulating a skeleton (or multiple connected skeletons if the distance between observations \( \Delta \) is large) is not a trivial task, especially in the more sophisticated versions of EA, and this must be repeated \( M \) times to obtain the Monte Carlo estimate of the transition density \( p(X_\Delta | X_0; \theta) \).
Chapter 4: Comparison of Estimation Procedures

We begin by comparing the methods of estimating the likelihood of discrete observations from an SDE by evaluating each method’s performance in terms of estimating the transition density. As the likelihood can be estimated by a straightforward combination of these individual estimated transition densities, the transition density has been the focus of the bulk of the existing literature.

4.1 Comparison of Estimates of the Transition Density

4.1.1 SMC vs. Hermite Expansions

Early comparisons of the approximate transition density based on SMC and the Hermite expansions [Aït-Sahalia, 2002a] were made in Aït-Sahalia [2002b] and in Jensen and Poulsen [2002] for univariate diffusion processes which are analytically reducible. That is, these processes can all be transformed to an SDE of unit diffusion using the Lamperti transform (2.13). Stramer and Yan [2007a] provide a comparison of the two methods for univariate diffusions which are only numerically reducible; that is, diffusions where

\[ \int \frac{du}{\sigma(u, \theta)} \]

does not exist in closed form and must be solved numerically. Overall, the general findings are that, when applicable, the Hermite expansion performs better from a

50
computational perspective. For small $\Delta$ values, the Hermite expansion proves more accurate than the estimates obtained via SMC. The converse is true for large values of $\Delta$ [Stramer and Yan, 2007a].

As the Hermite expansion is based on a Taylor series expansion near $\Delta = 0$, it requires that the time step between observations be very small. The threshold for “very small” will depend upon the functions $\mu(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ which govern the SDE. As a practical concern, assessing whether this condition has been met will be problematic. This is contrasted with SMC, where convergence may be investigated by comparing the estimates to those obtained for an increased $K$ and $M$ or by comparing estimates of the transition density for the same $K$ and $M$ across multiple runs of SMC.

While the estimated transition densities obtained from SMC are stochastic, those based on the Hermite expansion are deterministic. This difference means that, assuming that the practitioner has faith in the accuracy of the Hermite approximation, differentiation of the transition density and thus optimization to find the maximum of the approximated log-likelihood is relatively easy. On the other hand, differentiation of the log-likelihood estimated using SMC with respect to $\theta$ proves difficult due to the randomness inherent in these estimates. We address this shortcoming of SMC in the chapter that follows.

Also, while transformation of the SDE will often improve the performance of SMC [Stramer and Yan, 2007a], it is required for the Hermite expansion, which presents a problem when this transformation is not available. Additionally, Bakshi and Ju [2005] point out that while the Hermite expansion should, in theory, work for any one-dimensional SDE, the method not only assumes that the Lamperti transform can
be done analytically, but also that the inverse is known in exact closed form. The authors then provide two examples of SDEs which demonstrate this problem:

1. The short interest-rate model from Aït-Sahalia [1996] given by

\[ dX(t) = \mu(X(t), \theta)X(t)dt + \sqrt{\beta_0 + \beta_1 X(u) + \beta_2 X(u)^{\beta_3}}dW(t), \]

where \( X(0) = X_0 \) is the initial value of the interest rate, and \( \beta_0, \beta_1, \beta_2, \beta_3 \in \mathbb{R} \). Note that the necessary integral for the Lamperti transformation (2.13)

\[ \gamma(S(t); \theta) = \int^{s(t)} du \frac{du}{\sqrt{\beta_0 + \beta_1 X(t) + \beta_2 X(t)^{\beta_3}}}, \]

cannot be calculated analytically.

2. A stochastic volatility model for the log of the price of a stock \( X(t) \) given by

\[ dX(t) = \mu(X(t), \theta)dt + (\beta_0 + \beta_1 e^{-\beta_2 X(t)})dW(t), \]

where \( X(0) = X_0 \), is the log of the initial price of the stock and \( \beta_0, \beta_1, \beta_3 > 0 \). Note that the Lamperti transformation (2.13) can be calculated as

\[ \gamma(X(t); \theta) = \int^{X(t)} du \frac{du}{\beta_0 + \beta_1 e^{-\beta_2 u}} = \frac{X(t)}{\beta_0} + \frac{1}{\beta_0 \beta_2} \log \left( \beta_0 + \beta_1 e^{-\beta_2 X(t)} \right), \]

but this function is not invertible in closed form.

We next highlight an important property of the two estimation methods which will be investigated further in the context of the log-likelihood.

**Lemma 4.1.1.** For the SMC-based estimate of the transition density \( p^{(K,M)}(X(\Delta)|X_0, \theta) \), we have that

\[ p^{(K,M)}(X(\Delta)|X_0, \theta) > 0. \]
Proof. We have that
\[ p^{(K,M)}(X(\Delta)|X_0, \theta) = \frac{1}{M} \sum_{m=1}^{M} R_m^{(K)}, \]
so it is sufficient to show that \( R_m^{(K)} > 0 \) for \( m = 1, \ldots, M \). We have that
\[ R_m^{(K)} = \frac{\prod_{k=1}^{K} \xi(X_m(\tau_k)|X_m(\tau_{k-1}), \theta)}{q(X_m(\tau))} > 0, \]
since the numerator is the product of evaluations of a normal probability density function (pdf), each of which is strictly positive, and the denominator is the importance sampling density evaluated at a set of points drawn from this density, which must be strictly positive as well. Since the estimated likelihood is simply the product of these positive transition densities, it is also greater than zero.

The same cannot be said for the Hermite approximation (defined in (3.10) in terms of \( Z = \Delta^{-1/2} (Y(\Delta) - Y_0) \)) as originally presented in Aït-Sahalia [2002a], given by
\[ p_y^{(J)}(Y(\Delta)|Y_0, \theta) = \Delta^{-1/2} \phi\left(\Delta^{-1/2}(Y(\Delta) - Y_0)\right) \times \sum_{j=0}^{J} \eta_j(\Delta, Y_0; \theta) H_j \left(\Delta^{-1/2}(Y(\Delta) - Y_0)\right), \]
where the approximation is written in terms of the process \( Y \), obtained by transforming the process \( X \) using the Lamperti transform (2.13). The Hermite coefficients from (3.12) (expressed in terms of the process \( Y \)),
\[ \eta_j(\Delta, Y_0; \theta) = \frac{1}{j!} \mathbb{E}\left[H_j \left(\Delta^{-1/2}(Y(\Delta) - Y_0)\right) | Y_0, \theta\right], \]
are then approximated by a Taylor expansion of order \( \kappa \) (chosen to be 3 in the paper) in \( \Delta \). This truncated series can be negative, however, which gives a negative estimate of the transition density, and therefore a negative estimate of the likelihood. Section
2.2 of Stramer and Yan [2007a] and Section 4.2 Aït-Sahalia [2008] suggest addressing this issue by instead considering a Taylor series expansion of

$$\log \left( \sum_{j=0}^{J} \eta_j(\Delta, Y_0; \theta) H_j \left( \Delta^{-1/2}(Y(\Delta) - Y_0) \right) \right).$$

The Taylor coefficients may be obtained through the process of solving the Kolmogorov equations, expressing the equations in terms of $\Delta$ and the coefficients, and then matching the coefficients of $\Delta^k$, for $k = -1, 0, \ldots, \kappa$. They are no longer generally available in closed form, however, which is often thought to be the primary advantage of using the Hermite approximation to the transition density. The coefficients (denoted by $C$ to distinguish them from $\eta$) are given in Section 4.1 of Aït-Sahalia [2008] as

$$C_Y^{(-1)}(Y|Y_0) = -\frac{1}{2} (Y - Y_0)^2,$$

$$C_Y^{(0)}(Y|Y_0) = (Y - Y_0) \int_{\mu(Y_0 + u(Y - Y_0), \theta)}^{1} du,$$

$$C_Y^{(k)}(Y|Y_0) = k \int \limits_{0}^{1} G_Y^{(k)}(Y_0 + u(Y - Y_0)|Y_0) u^{k-1} du,$$

for $k \geq 1$, where

$$G_Y^{(1)}(Y|Y_0) = -\frac{d\mu(Y, \theta)}{dY} + \mu(Y, \theta) \frac{dC_Y^{(0)}(Y|Y_0)}{dY},$$

$$+ \frac{1}{2} \left( \left. \frac{d^2 C_Y^{(0)}(Y|Y_0)}{dY^2} \right|_{Y_0} + \left. \left( \frac{dC_Y^{(0)}(Y|Y_0)}{dY} \right)^2 \right|_{Y_0} \right)$$

and

$$G_Y^{(k)}(Y|Y_0) = -\mu(Y, \theta) \frac{dC_Y^{(k-1)}(Y|Y_0)}{dY} + \frac{1}{2} \frac{d^2 C_Y^{(k-1)}(Y|Y_0)}{dY^2},$$

$$+ \frac{1}{2} \sum_{h=0}^{k-1} \binom{k-1}{h} \frac{dC_Y^{(h)}(Y|Y_0)}{dY} \frac{dC_Y^{(k-1-h)}(Y|Y_0)}{dY},$$

for $k \geq 2$. 

54
Although this somewhat involved workaround exists, the original non-log transformed Taylor series expansion in $\Delta$ is commonly implemented in software packages (for example the $sde$ R package Iacus [2009b] which we use in later sections) and thus would appear to be the default method that many practitioners use. We now investigate the occurrence of these negative estimated transition densities.

To illustrate this, we consider a simulation study based on the Ornstein-Uhlenbeck process (2.19). This represents the best-case scenario, as the transition density being approximated (by an expansion around a normal distribution) is already normal. We assume that the diffusion coefficient is one (if this is not the case, the Lamperti transformation (2.13) may be used to obtain a process with unit diffusion) so that the SDE is defined by

\[
dX(t) = (\theta_0 + \theta_1 X(t)) \, dt + dW(t), \quad 0 \leq t \leq T,
\]

where $X(0) = X_0$ is the initial value of the process.

As shown in Chapter 2, conditionally on $X(0) = X_0$, for $\Delta > 0$, $X(\Delta)$ is normally distributed with mean and variance given by

\[
E(X(\Delta)|X_0) = X_0 e^{\theta_1 \Delta} - \frac{\theta_0}{\theta_1} \left(1 - e^{\theta_1 \Delta}\right) \quad \text{and}
\]

\[
\text{Var}(X(\Delta)|X_0) = \frac{1}{2\theta_1^2} (e^{2\theta_1 \Delta} - 1).
\]

For the values of $\Delta = 0.01, 0.05, 0.10, 0.25$, and $0.50$, we simulate $X_0$ from the stationary distribution of the OU process and then $X(\Delta)$ from the distribution conditional on $X_0$ with $\theta_0$ and $\theta_1$ selected to be 2 and -3, respectively. We then fix $\theta_1$ at the true parameter value used to simulate the $X$ values and use the Hermite approximation (available from the HPlglik function in the $sde$ R package [Iacus, 2009b]) to estimate the transition density over a range of $\theta_0$ values. We then record whether or
Table 4.1: Number (out of 100000 repetitions) of estimated likelihoods where a negative estimate of the transition density was obtained for at least one $\theta_0$ value using the Hermite approximation [Aït-Sahalia, 2002a] for various time steps $\Delta$. The initial and terminal observations were simulated exactly from the stationary and conditional distribution of (2.19), respectively. The results presented are for true parameter values of $\theta_0 = 2$ and $\theta_1 = -3$.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$\theta_0 \in [0, 5]$</th>
<th>$\theta_0 \in [1, 3]$</th>
<th>$\theta_0 \in [1.75, 2.25]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0.10</td>
<td>419</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>0.25</td>
<td>6263</td>
<td>231</td>
<td>15</td>
</tr>
<tr>
<td>0.50</td>
<td>4918</td>
<td>1713</td>
<td>1668</td>
</tr>
</tbody>
</table>

not there was a negative estimate of the likelihood for at least one of the considered $\theta_0$ values. Table 4.1 presents the number of negative estimates observed based on 100000 repetitions for the various $\Delta$ values. The columns represent different ranges of $\theta_0$ values where the estimates are obtained.

It has been previously noted [Stramer and Yan, 2007a] that the Hermite approximation is more accurate for (1) smaller values of $\Delta$ and (2) values of $\theta_0$ near the true parameter value. This same general statement appears to apply in terms of whether or not approximated transition densities are valid probability densities. This is evidenced in Table 4.1 by noting that the occurrence of negative probability estimates increases as $\Delta$ increases (moving downwards in the rows of the table) and as the range of $\theta_0$ candidate values narrows (moving leftward in the columns of the table). The frequency of these negative values is striking. For example, for $\Delta = 0.25$, when the Hermite expansion is used to estimate the transition density at the true parameter value and $\theta_1 = -3$ and over the range of $[0,5]$ for $\theta_0$ for the OU model where the transition density is already normal, at least one estimated transition density is not
a valid probability roughly 6% of the time (with an estimated standard error of approximately 0.08%). For higher dimensional problems (e.g., we have at least $\theta_0$ and $\theta_1$ values), we have found that the occurrence of negative estimates will increase (as a larger proportion of the space considered contains parameter values of low likelihood). We highlight here that these results represent only a single transition density; that is, a pair of points $X_0$ and $X_\Delta$. In the section that follows, we investigate the overall effect on the estimated log-likelihood, as we only need one negative density out of the $N$ overall transition densities to cause problems in estimating the log-likelihood.

4.1.2 SMC vs. EA

The estimated transition density based on the exact algorithm (EA) can be thought of as a special case of SMC. In particular, it is the limiting (as $K$ goes to infinity) case of Pedersen’s estimate of the transition density. As mentioned earlier, the optimal choice of $q(\cdot)$ is the true joint density of $X(\tau)$ given $X_0$ and $X(\Delta)$ [Stramer and Yan, 2007a], which is unavailable, as it depends on the transition density that we are trying to estimate. However, using EA we are able to sample from $q(\cdot)$ that is the true joint density of $X(\tau)$ given $X_0$. That is, obtaining estimates using EA can be thought of as removing the discretization error inherent in Pedersen’s sampler, which relies on the Euler approximation of the process. Alternatively, when estimates are obtained using the sampler of Durham and Gallant, the discretization error remains, but the variates $X(\tau)$ are generated conditionally on $X(\Delta)$ in addition to $X_0$.  

57
In the SMC context the estimate of the transition density based on EA can be written as

\[ p^{(M)}(X(\Delta)|X_0, \theta) = \frac{1}{M} \sum_{m=1}^{M} p(X_{\Delta}|X_0; \theta, S), \]

where

\[ S = ((0, X_0), (t_1, X^*_1), \ldots (t_k, X^*_k), (T, X_T)), \]

is the skeleton generated using EA on \([0, T]\), where \(T > \Delta\). The exact form of \(p(X_{\Delta}|X_0; \theta, S)\) will depend on which version of EA was used to simulate the skeleton \(S\). For example, in EA1 the exact density, conditional on the skeleton is the law of a Brownian bridge conditioned on the two points of the skeleton immediately on either side of \(X(\Delta)\). Note that in EA the superscript \(K\) is absent from the transition density, as the discretization error has been removed by the use of the skeleton. That is, the only source of error is the Monte Carlo variability. As \(K \to \infty\), Bally et al. [1995] show that

\[ p^{(K)}(X(\Delta)|X_0, \theta) \to p(X(\Delta)|X_0, \theta), \]

so that the integral based on the auxiliary \(X(\tau)\) values used to compute \(p^{(K)}(X(\Delta)|X_0, \theta)\) becomes closer to the true transition density as \(K\) becomes larger. This reduction in bias is demonstrated in Figure 4.1, which shows the histograms of 10,000 simulations from the Euler scheme for various \(K\) values with the true OU transition density overlaid (again using the true parameter values \(\theta_0 = 2\) and \(\theta_1 = -3\)). Here \(X_0 = 0.968\) was simulated from the stationary distribution of (2.19) and we consider the approximation to the transition density at \(X(1)\). These plots visually demonstrate the bias due to simulation decreasing as \(K\) increases. We note that, even for fairly large values of \(K\), the bias does not appear to completely disappear.
Figure 4.1: Density plots of realized Monte Carlo samples of the Euler approximation to the Ornstein-Uhlenbeck transition density for various numbers of fill-in points, $K$ (blue lines). The true parameter values for the process are $\theta_0 = 2$ and $\theta_1 = -3$. The black lines represent the true transition density of $X(1)$ given $X_0 = 0.968$, from the conditional distribution of (2.19).
Originally, Durham and Gallant [2002] considered improvements to Pedersen’s sampler in two respects: (1) bias reduction (reducing discretization error) and (2) variance reduction (reducing the Monte Carlo error). The first error arises due to the approximation of the continuous-time process with a discrete-time process, while the second arises due to the evaluation of the transition densities using simulation. The first error can be improved by considering higher order (and more computationally intensive) schemes than the Euler approximation employed in Pedersen’s sampler. The second error is the more problematic of the two, as Pedersen’s sampler ignores the endpoint, and thus generates many paths which are known a priori to be very unlikely. By conditioning on $X(\Delta)$ in addition to $X_0$, the sampled paths $X(\tau)$ come from regions of much higher density and the improvement is “truly amazing” [Brandt and Santa-Clara, 2002a]. Although Durham and Gallant [2002] consider discuss methods to reduce both types of error, it has been generally remarked [Aït-Sahalia, 2002b, Brandt and Santa-Clara, 2002a] that the most important contribution to improving the approximation is the variance reduction in the estimated transition density. That is, the biggest problem with using Pedersen’s method to approximate the transition density is due to the fact that the endpoint $X(\Delta)$ may be far away from the simulated values rather than the discretization error inherent in the Euler approximation. The exact methods essentially solve the latter problem but ignore the more important former problem. This problem is demonstrated visually in Figure 4.2, which is based on Figure 3 of Durham and Gallant [2002]. Using a sampler based only on $X_0$, the random variates may be far away from $X(\Delta)$, thus the samples are most often drawn from regions where the integrand in the discretized likelihood (3.15) has little mass. This demonstrated in the top (Pedersen’s sampler) plot of Figure 4.2, as the black
lines, which indicate the distance from the observed point to the simulated points used to approximate the transition density. In the bottom plot, due to structure imposed on the simulated values using Durham’s and Gallant’s modified Brownian bridge sampler, the black lines are much shorter, as the simulated points are much closer to the observed point.

We also note here some disadvantages with the EA approaches. Many commonly used models will not fit into the framework of EA1 or EA2, and will require the simulation of layered Brownian bridges involved in EA3, which is a non-trivial computation. Note also that using EA, the time points at which the skeletons are observed is not known beforehand, in contrast to the usual SMC estimation procedures. Thus, the simulation of multiple skeletons may be required to ensure that points are simulated near the time ∆, which can further add to the computational cost.

4.2 Comparison of Estimates of the Likelihood

In the previous section, we discussed the performance of the different schemes for approximating the transition density. We now examine their performance in terms of approximating the log-likelihood, the combination of \( N \) estimated transition densities. We find that the approximations based on both SMC and the Hermite expansion encounter difficulty in the tails of the distribution. As our investigation will demonstrate, this problem is easily solved in the SMC case by simply increasing the number of Monte Carlo samples. While the Hermite expansion based on higher order terms may improve performance, this will necessarily be more involved than simply changing one of the inputs to SMC. For example, as we stated before, the Hermite expansion
Figure 4.2: Approximate simulated variates based on only $X_0$ (Pedersen) and approximate simulated variates based on both $X_0$ and $X(\Delta)$ (Durham and Gallant) for $\Delta = 1$. The gray lines indicate simulated $X(\tau)$ values and the black lines indicate the calculation of the Monte Carlo variate.
of the (non-log) density is typically implemented in software packages using 6 Hermite polynomials with coefficients approximated using a Taylor series approximation of order 3. Including higher order terms will involve derivation and implementation steps that will not be necessary for SMC.

We previously investigated the accuracy of the estimated transition densities based on the Hermite expansion. We now turn to the problem of estimating the log-likelihood of data for a given model and set of parameter values. To investigate the effect that the sampling interval and the number of data points has on these estimates, we consider various specifications of each. As in the examination of the approximate transition densities, we use simulated data from the OU model (2.19) with true parameter values \( \theta_0 = 2 \) and \( \theta_1 = -3 \). The Hermite approximation is then computed using the \texttt{HPloglik} function in the \texttt{sde} R package [Iacus, 2009b] over 100 possible \( \theta_0 \) values in \([0, 5]\). In Figure 4.3, the estimates of the log-likelihood are presented graphically for a single set of simulated observations at each combination of the number of data points \( N \) and the sampling interval \( \Delta \). As in the case of the individual transition densities, we see that as \( \Delta \) increases (moving rightward in the columns of Figure 4.3), the Hermite expansion becomes problematic. We note that the places where no red line is visible correspond to a negative estimate of the likelihood and thus an undefined value for the log-likelihood. We found earlier that the Hermite expansion ran into trouble in the tails of the distribution, which explains why the estimates are available less often as the number of observed data points \( N \) increases (moving downward in the rows of Figure 4.3). The plots in Figure 4.3 demonstrate the comparison of the true log-likelihood and the Hermite approximation for a single set of simulated values in each of the 12 cases. We then repeat this for 1000 simulated
Table 4.2: Mean (standard deviation) of the observed proportion (based on 1000 simulated datasets each) of negative estimated likelihoods using the Hermite approximation. The estimated likelihoods were calculated for 100 possible values of $\theta_0 \in [0, 5]$, where the true parameter value of $\theta_0 = 2$ and $\theta_1$ is fixed at the true value -3. These quantities are computed for various time steps $\Delta$ and various numbers of observed data points $N$. The estimated standard errors for these estimates were at most 0.0015.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>500</td>
<td>0.002 (0.015)</td>
</tr>
<tr>
<td>1000</td>
<td>0.003 (0.020)</td>
</tr>
<tr>
<td>5000</td>
<td>0.023 (0.053)</td>
</tr>
<tr>
<td>10000</td>
<td>0.039 (0.070)</td>
</tr>
</tbody>
</table>

datasets in each of the 12 cases and investigate how often the Hermite approximation of the log-likelihood is unavailable. We then record the proportion of these 100 estimates which are undefined. The mean and standard deviation (in parentheses) of these proportions based on 1000 simulations in each of the 12 cases is presented in Table 4.2. Again we see that as $N$ and $\Delta$ increase, the proportion of negative estimates of the likelihood increases and that these proportions are often strikingly large. Note that, as when we evaluated the estimated transition densities, this model represents the best-case scenario, as the transition density being approximated (by an expansion around a normal distribution) is already normal. For more complicated processes, the performance of the Hermite estimates could begin to suffer for even smaller values of $N$ and $\Delta$. As in the estimation of the transition density, the Hermite approximation to the log-likelihood, when available, will have very little bias and zero variance. However, the availability of these estimates will be a much larger concern in the case of the log-likelihood as the number of observations grows.
Figure 4.3: True log-likelihoods for $N$ data values simulated from an OU model with $\Delta$ time between observations and true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ (black lines) and corresponding estimated log-likelihoods based on the Hermite expansion of Aït-Sahalia [2002a] (red lines).
Figure 4.4: True log-likelihoods for $N$ data values simulated from an OU model with $\Delta$ time between observations and true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ (black lines) and corresponding estimated log-likelihoods based on the SMC scheme of Pedersen [1995a] with $K = 5$ subintervals and $M = 25$ Monte Carlo samples (red lines).
The SMC method can encounter difficulties caused by the tail of the distribution as well. However, as we will demonstrate, these difficulties can be solved by a simple change to the numeric values input to SMC and some additional computation time. We also have found that the usual guidance on choosing these inputs for estimating the transition density can be misguided for estimating the log-likelihood. To investigate this, we simulate data from the OU model (2.19) with $\theta_0 = 2$ and $\theta_1 = -3$, separated in time by increments of $\Delta = 0.05, 0.1, \text{and} 0.25$. We then fix $\theta_1 = -3$ and examine the true and estimated log-likelihoods as a function of $\theta_0$, as in the Hermite expansion case, for various $K$ and $M$ values.

The contrast between the Hermite approximation and the SMC-based estimates is demonstrated visually in Figure 4.4. We provide this figure for ease of comparison to Figure 4.3 and due to the fact that the scale of the true and estimated log-likelihoods (and thus the corresponding errors) varies widely. As we are most interested in the error, or the difference between the true log-likelihood and the estimated log-likelihood obtained via SMC, we will focus on this difference for the remainder of this section. For this, we turn to Figure 4.5, which shows 16 panels corresponding to those of Figure 4.4, but in terms of the errors (displayed on a common scale).

As previously discussed, the estimated likelihoods using SMC are guaranteed to be positive, so there are no issues with the availability of estimated log-likelihoods as there were in the Hermite expansion case. The sampling variability in the estimates based on SMC is evident as well, as opposed to the fixed deterministic nature of the Hermite expansion. (This sampling variability can cause difficulty in maximizing the log-likelihood, an issue we address in the chapter that follows.) There does not appear to be the deterioration in the performance of SMC as $\Delta$ and $N$ grow large that was
Figure 4.5: Difference between true log-likelihoods and estimated log-likelihoods for $N$ data values simulated from an OU model with $\Delta$ time between observations and true parameter values $\theta_0 = 2$ and $\theta_1 = -3$. The SMC-based estimates are obtained based on the Pedersen’s scheme with $K = 5$ subintervals and $M = 25$ Monte Carlo samples.
evident in the Hermite expansion. While the estimates are always available via SMC, there is a downward bias in the estimates compared to the true log-likelihood for this particular choice of 5 subintervals \((K)\) and 25 Monte Carlo samples \((M)\). We now examine the cause of this bias and how to adjust the SMC inputs \(K\) and \(M\) to remedy it.

A common claim in the existing literature is that increasing \(K\) will decrease the discretization error (bias) and increase the Monte Carlo error (variability) while increasing \(M\) will reduce the variability. The authors of Brandt and Santa-Clara [2002b] claim “Increasing \([K]\) or \([M]\), or both, improves the approximation of the transition densities and thus results in an estimator that behaves more like the exact MLE”, those in Durham and Gallant [2002] state “Increasing \([K]\) reduces bias, but at the cost of greater variance”, to list a few. Visually, this is illustrated in Figure 4.6, where we observe that increasing \(K\) appears to increase Monte Carlo variability. This variability dominates the plot when \(M = 100\), so it is hard to infer much about the bias in that setting. However, when \(M\) is increased to 10000, we are able to see the decrease in the bias of the estimator when we move from \(K = 2\) to a larger \(K\), although it is difficult to visually distinguish between \(K = 5, 10,\) and 50, except for the increased variability as \(K\) increases.

These results generally agree with what is reported in the literature. That is, increasing \(K\) reduces bias at the cost of increased variability and increasing \(M\) reduces variability. This is not surprising, as the existing literature often considers the transition density \((N = 1)\) rather than the log-likelihood based on a large \(N\).

However, we have found that for any \(M\), increasing \(K\) past some optimal choice actually *increases* the bias in the estimated likelihood. In the following section, we will
Figure 4.6: Difference between the true log-likelihood and estimated log-likelihood obtained via SMC for $N = 2$ observations observed $\Delta = 0.1$ apart in time from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ as a function of $\theta_0$ for different numbers of fill-in points between observations $K$ and numbers of Monte Carlo samples $M$. The horizontal red line corresponds to no error.
show that the typical statement that increasing $K$ reduces the bias of the estimated log-likelihood may be true in theory, but in practice, increasing $K$ for a fixed $M$ may actually cause us to obtain a worse estimate of the likelihood at the expense of extra computational cost. This problem grows worse as $N$, the amount of data, increases.

4.2.1 Effect of $K$ on Bias of SMC Estimates

While it is true that the difference between $p^{(K)}(X(\Delta)|X_0, \theta)$ (that is, $p^{(K,M)}(X(\Delta)|X_0, \theta)$ with $M = \infty$) and the true transition density $p(X(\Delta)|X_0, \theta)$ goes to zero as $K$ increases, we now graphically illustrate what can happen as $K$ increases for a given $M$. In Figure 4.7, we plot errors in the SMC estimates of the log-likelihood of $N = 1000$ observations observed $\Delta = 0.1$ apart in time from the OU model (2.19) with $\theta_0 = 2$ and $\theta_1 = -3$. In the earlier plots, we considered the effect of $N$ and $\Delta$, things that are not within the practitioner’s control, but we now consider varying $M$ and $K$, which may be chosen or adjusted at will in a given setting. We notice something unexpected happening for $M = 100$. Here, increasing $K$ actually increases the bias in the estimate of the log-likelihood. That is, as the estimates become more computationally expensive to obtain, they also become more inaccurate. This phenomenon does not immediately occur for $M = 10000$ however, and only begins to show up when $K$ is increased from 10 to 50. The mean and variance (over $\theta_0$) of the errors are displayed in Table 4.3. The numbers in this table support what we observe visually in Figure 4.7. That is, increasing $K$ increases the variability of the estimates regardless of $M$, and for $M = 100$, increasing $K$ actually increases the bias in the estimates. On the other hand, when $M = 10000$, increasing
Table 4.3: Mean and variance (over $\theta_0$) of the difference between the true log-likelihood and the estimated log-likelihood for $N = 1000$ observations observed $\Delta = 0.1$ apart in time from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ for different numbers of fill-in points between observations $K$ and numbers of Monte Carlo samples $M$.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$M = 100$</th>
<th>$M = 10000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
</tr>
<tr>
<td>2</td>
<td>8.73</td>
<td>4.83</td>
</tr>
<tr>
<td>5</td>
<td>15.27</td>
<td>62.50</td>
</tr>
<tr>
<td>10</td>
<td>33.86</td>
<td>348.80</td>
</tr>
<tr>
<td>50</td>
<td>191.02</td>
<td>11858.42</td>
</tr>
</tbody>
</table>

$K$ up to some point reduces the average bias, but increasing $K$ further appears to increase the bias.

To investigate where this discrepancy between small and large $N$ is coming from, we separate the log-likelihood into the 1000 individual log transition densities. In Figure 4.8, we have the individual contributions to the log-likelihood for each of the 1000 pairs of points. The log-likelihood contribution of $p(X(46.5) = -0.596|X(46.4) = 0.185)$ (highlighted) appears to be different from the others for many of the $\theta_0$ values considered.

We see from Figure 4.9 that in the $K=50$ case, there is a substantial bias for many of the $\theta_0$ values on the order of 50 or 100 on the log-likelihood scale, which accounts for a substantial portion of the difference seen in Figure 4.7. We compare this with another (arbitrarily chosen) pair of points, $X(40)$ and $X(39.9)$ to appreciate the scale of this bias relative to the well-behaved points. Although there is variability in the estimates of the log transition density of $X(40)$ for increasing $K$ values, it is masked by the scale of the plot, specified to be the same for both points. The mean and
Figure 4.7: Difference between true log-likelihood and estimated log-likelihood obtained via SMC for $N = 1000$ observations observed $\Delta = 0.1$ apart in time from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ as a function of $\theta_0$ for different numbers of fill-in points between observations $K$ and numbers of Monte Carlo samples $M$. The horizontal red line corresponds to no error. We note the very different scales of the left and right columns, which correspond to $M = 100$ and $M = 10000$ Monte Carlo samples, respectively.
Figure 4.8: Individual log transition densities and overall log-likelihood of $N = 1000$ observations spaced $\Delta = 0.1$ in time from an OU model with $\theta_0 = 2$ and $\theta_1 = -3$. 
Table 4.4: Mean and variance (over $\theta_0$) of the difference between the true and SMC-based estimates of log transition density for extreme value $X(46.5)$ and non-extreme value $X(40)$ for various numbers of subintervals $K$ and $M = 100$ Monte Carlo samples.

<table>
<thead>
<tr>
<th>$K$</th>
<th>Extreme Mean</th>
<th>Variance</th>
<th>Non-extreme Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.43</td>
<td>0.26</td>
<td>0.05</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>1.05</td>
<td>3.40</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>10</td>
<td>3.21</td>
<td>16.40</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>20.90</td>
<td>445.40</td>
<td>0.04</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Variance (over $\theta_0$) of the bias in the estimated transition densities corresponding to the extreme value and the non-extreme value is displayed in Table 4.4. The numbers in this table support what we observe visually in Figure 4.8. That is, the mean and the variance of the bias are much larger for the extreme observations and increasing $K$ while $M$ remains fixed seems to further increase these quantities.

We now consider the behavior of the estimated log-likelihood based on $X(46.5)$ and $X(46.4)$. The left hand plot in Figure 4.10 shows 10000 realizations from the exact distribution of the estimator at the true parameter values $\theta_0 = 2$ and $\theta_1 = -3$, with the blue and black (overlapping since the bias is so small) lines indicating the log of estimated likelihood based on these 10000 realizations and the true log-likelihood. Additionally, the same histogram is shown on the likelihood scale, with x symbols drawn below the x axis at those likelihoods which were observed in the 10,000 realizations. The extreme skew should be noted in the top half Figure 4.10, as only 0.23% of the observations are larger than the true mean. That is, although the theoretical mean of the SMC-based estimates of the log-likelihood is nearly identical to the true log-likelihood, the distribution is so skewed that for a smaller Monte Carlo
Figure 4.9: Difference between true and estimated log transition densities obtained via SMC for extreme value $X(46.5)$ and non-extreme value $X(40)$ for various numbers of subintervals $K$ and $M = 100$ Monte Carlo samples. The horizontal red line corresponds to no error. We note the very different scales of the left and right columns, which correspond to the extreme value and non-extreme value, respectively.
sample of these estimates, it will often be the case that the estimated log-likelihood is heavily biased. For comparison, identical plots are created based on $X(40)$ and $X(39.9)$ in the bottom half of Figure 4.10 (the x symbols marking the locations of the observations show up as a black line due to the fact that these observations are more spread out. Although this distribution of this estimate is still skewed, it is clearly much less skewed than the distribution of the estimated transition density for the extreme value.

In our examination of the Hermite expansion, we found that the estimated log-likelihood would often be undefined, often due to only a handful of observations. That is, of the $N$ log transition densities summed to obtain the log-likelihood, only one log transition density is required to be undefined to cause problems in the overall log-likelihood estimate. These problems were often caused by “extreme” observations, pairs of sequential observations with very low conditional probabilities. Although the estimates based on SMC do not encounter the problem of being undefined, as seen in Figure 4.7, a downward bias in the estimates may appear. As indicated by Figure 4.9, it appears that, as in the case of the Hermite approximation, the bulk of the problem is being caused by a handful of extreme observations.

4.2.2 Investigating the Source of the Bias Further

We now examine the cause of this downward shift in the estimated log-likelihood (which leads to larger errors) obtained via SMC by considering the distribution used to generate these estimates. We note that depending on the specific context, this bias may or not be problematic. As the shape of the log-likelihood is typically preserved,
Figure 4.10: Sampling distribution of log transition density and transition density estimates for extreme value $X(46.5)$ (top) and non-extreme value $X(40)$ (bottom) evaluated at true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ using $K = 50$ subintervals. The black x symbols below the x-axis indicate which values of the likelihood were observed. In the non-extreme case, the observations are less skewed, and these x symbols create a solid black line.
maximum likelihood estimation may not be affected. However, when the actual evaluation of the log-likelihood is important, this bias should not be ignored. In Brandt and Santa-Clara [2002b], the convergence of the simulated MLE to the true MLE is shown when $K \to \infty$, $M \to \infty$, and $\sqrt{M}/K \to 0$. The true MLE in turn converges to the true parameter vector as $N \to \infty$. As suggested visually in Figure 4.7 and as the following section will demonstrate, the likelihood is most biased when $\sqrt{M}/K$ is small and $N$ is large (which may or may not be reflected in the simulated MLE).

The situation appears to be a bit more complex than simply increasing $K$ to reduce bias as often recommended in the existing literature, which we illustrate for Pedersen’s sampler. We can think of dividing the distance between observations into $K$ equally spaced intervals of length $\Delta/K$ as a special case of a more general setting. Namely, that the distance is divided into $K = K_1 + 1$ intervals, where the first $K_1$ intervals are of equal length

$$\frac{\Delta}{K_1} \left( \frac{K_2 - 1}{K_2} \right)$$

similar to before, but scaled so that the final interval ends at time

$$\Delta \left( 1 - \frac{1}{K_2} \right).$$

(There is no requirement that these $K_1$ intervals be of equal length, but there is no apparent benefit to varying these lengths to our knowledge.) The final interval, between the last simulated point and the second of the pair of observed values, is of length $\Delta^* \equiv \Delta/K_2$.

For notational simplicity, assume that we observe $X(0) = X_0$ and $X(1)$ and that we wish to approximate the transition density between the two. Recalling the simulation step of SMC (3.18), for each $m = 1, \ldots, M$ we simulate $K_1$ fill-in points
until we have $X_m (1 - \Delta^*)$. The approximation of the likelihood from $X_m (1 - \Delta^*)$ to $X(1)$ is a random variable, so we first examine its distribution, which can then be used to investigate what happens for the average of $M$ of these IID RVs.

Once we understand the effect of $K_1$, which controls the level of discretization error in the simulated fill-in variates, and accept the fact that with unlimited computation, we would choose $K_1 = \infty$ (and thus be simulating these variates exactly), we can examine what happens for different values of $K_2$ when $K_1 = \infty$ for the OU process, as in this case we have available the true exact conditional distribution of $X_m (1 - \Delta^*)|X_0$. Recall from the definition of the OU process (2.19) that the true conditional density of $X_m (1 - \Delta^*)$ given $X_0$ is Gaussian with

$$E[X_m (1 - \Delta^*) | X_0] = X_0 e^{\theta_1 (1 - \Delta^*)} - \frac{\theta_0}{\theta_1} (1 - e^{\theta_1 (1 - \Delta^*)})$$ and

$$\text{Var}[X_m (1 - \Delta^*) | X_0] = -\frac{\gamma^2}{2\theta_1} (1 - e^{2\theta_1 (1 - \Delta^*)}),$$

so we simulate $X_m (1 - \Delta^*)$ from this distribution and, then calculate the Euler approximation to the transition density of $X(1)$ given $X_m (1 - \Delta^*)$, which we denote by $L^*$. We have

$$L^* = \frac{1}{\sqrt{2\pi \gamma^2 \Delta^*}} \exp \left\{ \frac{(X(1) - X_m (1 - \Delta^*) - (\theta_0 + \theta_1 X_m (1 - \Delta^*) \Delta^*)^2}{-2\gamma^2 \Delta^*} \right\},$$ (4.1)

since $L^*$ is the density of a normal random variable with mean

$$X_m (1 - \Delta^*) + (\theta_0 + \theta_1 X_m (1 - \Delta^*)) \Delta^*,$$

and variance $\gamma^2 \Delta^*$ evaluated at $X(1)$.

To find the distribution of $L^*|X_0$, we begin by noting that

$$X(1) - X_m (1 - \Delta^*) - (\theta_0 + \theta_1 X_m (1 - \Delta^*) \Delta^*$$

$$= X(1) - (1 + \theta_1 \Delta^*) X_m (1 - \Delta^*) - \theta_0 \Delta^*$$ (4.2)
follows a normal distribution with mean $\mu_L$ and variance $V_L$, where

$$\mu_L = X(1) - \theta_0 \Delta^* - (1 + \theta_1 \Delta^*) \left[ X_0 e^{\theta_1(1-\Delta^*)} - \frac{\theta_0}{\theta_1} (1 - e^{\theta_1(1-\Delta^*)}) \right]$$

and

$$V_L = -(1 + \theta_1 \Delta^*)^2 \frac{\gamma^2}{2\theta_1} \left(1 - e^{2\theta_1(1-\Delta^*)}\right).$$

We now use the fact that if a random variable $U$ follows a normal distribution with mean $m$ and variance $\sigma^2$ then

$$\frac{U^2}{\sigma^2} \sim \chi^2_1\left(\frac{m^2}{\sigma^2}\right)$$

where $\chi^2_k(\lambda)$ is the noncentral chi-square distribution with $k$ degrees of freedom and noncentrality parameter $\lambda$.

Using this result with the quantity in (4.2) in place of $U$ gives

$$\frac{(X(1) - (1 + \theta_1 \Delta^*)X_0 (1 - \Delta^*) - \theta_0 \Delta^*)^2}{V_L^2} \sim \chi^2_1(\lambda),$$

where the noncentrality parameter is given by

$$\lambda = \frac{\mu_L^2}{V_L}.$$

Using the distribution allows us to rewrite the approximation to the transition density (4.1) as

$$L^*|X_0 = \frac{1}{\sqrt{2\pi\gamma^2\Delta^*}} \exp(cY),$$

(4.3)

where $Y \sim \chi^2_1(\lambda)$ and

$$c = \frac{(1 + \theta_1 \Delta^*)^2 \gamma^2}{2\theta_1} \left(1 - e^{2\theta_1(1-\Delta^*)}\right).$$ 

For the noncentral chi-square distribution with 1 degree of freedom and noncentrality parameter $\lambda$, we have

$$f_Y(y) = \frac{1}{2} \exp\left\{-(y + \lambda)/2\right\} \frac{y^{1/2}}{\lambda} I_{-1/2}(\sqrt{\lambda y})$$

(4.4)
for \( y \geq 0 \), where \( I_k \) is the modified Bessel function of the first kind of order \( k \).

From András and Baricz [2008], the pdf of \( y \) can also be written in the following form, which may be easier to deal with:

\[
f_Y(y) = \frac{\exp\left\{-(y + \lambda)/2\right\}}{\sqrt{2\pi y}} \cosh(\sqrt{\lambda y})
= \frac{\exp\{- (y + \lambda)/2\}}{\sqrt{2\pi y}} \left( \exp\left\{\sqrt{\lambda y}\right\} + \exp\{- \sqrt{\lambda y}\} \right) / 2 \tag{4.5}
\]

In order to obtain the distribution of \( L^\ast | X_0 \), we invert (4.3) to obtain

\[
Y = \log\left\{\sqrt{2\pi \gamma^2 \Delta^\ast L^\ast} \right\}/c,
\]
with

\[
\frac{dY}{dL^\ast} = \frac{1}{cL^\ast}.
\]

These two pieces allow us to write

\[
f_{L^\ast}(L^\ast) = \frac{\exp\left\{- \left( \log\left\{\sqrt{2\pi \gamma^2 \Delta^\ast L^\ast} \right\}/c + \lambda \right) / 2 \right\}}{\sqrt{2\pi \log\left\{\sqrt{2\pi \gamma^2 \Delta^\ast L^\ast} \right\}/c}}
\times \left( \exp\left\{\lambda \log\left\{\sqrt{2\pi \gamma^2 \Delta^\ast L^\ast} \right\}/c \right\} + \exp\left\{- \lambda \log\left\{\sqrt{2\pi \gamma^2 \Delta^\ast L^\ast} \right\}/c \right\} \right) / 2 \times \frac{1}{cL^\ast} \right. \tag{4.6}
\]

We now investigate what happens as \( \Delta^\ast \to 0 \). Recall that in the usual setting, this is equivalent to \( K \to \infty \). Plugging \( \Delta^\ast = 0 \) for \( \mu_L \) and \( V_L \) yields

\[
\mu_L = X(1) - \left( X_0 e^{\theta_1} - \frac{\theta_0}{\theta_1} (1 - e^{\theta_1}) \right) \quad \text{and} \quad \nu_L = -\frac{\gamma^2}{2\theta_1} (1 - e^{2\theta_1}),
\]
so that

\[ \lambda = \left( X(1) - \left( X_0 e^{\theta_1} - \frac{\theta_2}{\theta_1} (1 - e^{\theta_1}) \right) \right)^2 \] and

\[ \lim_{\Delta^* \to 0} c = \lim_{\Delta^* \to 0} \frac{(1 + \theta_1 \Delta^*)^2 \frac{\gamma^2}{2\theta_1} (1 - e^{2\theta_1 (1-\Delta^*)})}{2\gamma^2 \Delta^*} = -\infty, \]

since

\[ (1 + \theta_1 \Delta^*)^2 > 0, \quad (1 - e^{2\theta_1 (1-\Delta^*)}) > 0, \quad \text{and} \quad \frac{\gamma^2}{2\theta_1} < 0. \]

To summarize, as the distance between the final simulated point and the ending observation, \( \Delta^* \), goes to zero, the distribution of \( Y \) converges to a noncentral chi-square distribution with one degree of freedom and noncentrality parameter given above. However, the estimates of the likelihood are \( Y \) multiplied by an increasingly large negative number (as \( \Delta^* \to 0 \)) and then exponentiated. This is the cause of the skewed distributions displayed in Figure 4.10. So, for a fixed \( M \), which can be thought of as the finite sample size from the true theoretical distribution (4.6), increasing the number of subintervals \( K \) increases the skew of this distribution and can actually increase the bias in the estimated log-likelihood, in contrast to what is often reported.

This effect is exaggerated as the number of data points increases, as observations in the tail of the conditional distribution become increasingly common. Note that when \( X(1) \) is in the tail of the distribution (that is, far from its mean based on the Euler approximation), the noncentrality parameter \( \lambda \) will become larger. This causes the distribution of the estimates to become increasingly skewed. Although observations in the tail of the distribution will become increasing prevalent as the number of observations \( N \) grows, this can be remedied by increasing the number of
Monte Carlo samples. That is, by sampling enough points to ensure that the tail of the distribution has been well represented in the sample, we obtain an unbiased estimate of the true transition density. We suggest that rather than using standard rules of thumb for choosing $M$ solely determined by $K$, the size of $N$ should factor into this choice as well. Although simple guidelines such as $M = K^2$ are easy to remember and appear to be optimal for estimating a single transition density, in practice, $N$ cannot be ignored. When the true transition density is unknown, the choice of $M$ should be made while keeping $N$ in mind, and perhaps repeating the estimation procedure for multiple $M$ values to check whether there appears to be any shift in the estimated log-likelihood. In the data analysis of Chapter 6 we compared the estimates obtained for a handful of $\theta$ values to those which would have been obtained for using twice as many Monte Carlo samples to avoid this downward bias.

To demonstrate how increasing $M$ can improve this bias, we turn to Figure 4.11. The three plots on the left side show the true log-likelihood (black line) and smoothed versions of the estimated log-likelihoods (to remove the sampling variability that often masks bias) using Pedersen’s sampler for a large number of different $K$ values. We note that (moving from top to bottom) as the number of Monte Carlo samples increases, the negative bias for the large $K$ values shrinks and the estimated log-likelihood becomes more accurate. That is, in the case of $M = 100$, the smaller $K$ values produce SMC-based estimates closer to the truth than the large $K$ values. This pattern is diminished, but still visible when $M = 1000$. When $M$ is increased to 10000, the pattern is reversed, and the smallest $K$ ($K = 2$) is no longer the choice which produces the least biased estimates. Note that here we have a moderate number of observations ($N = 1000$), but that this is large enough to affect the optimal choice.
of $K$ and $M$. For example, in the first plot, we see that $K = 2$ produces the most accurate estimates of the log-likelihood for $M = 100$ Monte Carlo samples, which disagrees with the usual suggestion $M = K^2$ for the transition density. Also we consider the mean (over $\theta_0$) absolute difference between the estimates and the true log-likelihood. These are plotted as a function of the number of fill-in points $K$ in the right hand side of Figure 4.11, where we observe the same pattern as in the plots on the left.

Also note that, although the problem may still occur using the modified Brownian bridge sampler of Durham and Gallant [2002], it will be less prevalent, at least in the OU case. This is because, by considering both the initial and terminal points when generating the fill-in values, the simulated variates are pulled toward the end point, so that $X(1)$ will be in the tail of the distribution much less often than when the variates were generated using Pedersen’s method. This is illustrated visually in Figure 4.12, where we display histograms of the estimated log-likelihoods for the extreme observation previously discussed. Again the sample mean of the 10000 observations is nearly identical to the true log-likelihood, but the shapes of these distributions look very different. To highlight this difference, we plot the histograms on two different scales.

We emphasize that although we choose to use the modified Brownian bridge sampler of Durham and Gallant [2002], in the remainder of this dissertation, the user is free to choose any $q(\cdot)$ desired before proceeding with our proposed method. As importance samplers continue to be developed to improve the performance of SMC in various settings, these can be plugged directly into the optimization scheme that follows. This chapter has provided guidance for obtaining estimates at a fixed $\theta$,
Figure 4.11: Smoothed estimates obtained using Pedersen’s sampler for various combinations of the number of data points $N$, the number of fill-in points $K$, and the number of Monte Carlo samples $M$. The data were obtained from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$ with observations spaced $\Delta = 0.1$ in time.
Figure 4.12: Sampling distribution of log transition density estimates for extreme value \( X(46.5) \) evaluated at true parameter values \( \theta_0 = 2 \) and \( \theta_1 = -3 \) using \( K = 50 \) subintervals obtained using the sampler of Durham and Gallant (left) and Pedersen’s sampler (right). The top row of histograms are drawn at their respective default scales, while the bottom row is plotted on the same scale to highlight the skew of the distribution of Pedersen’s estimates.
but exploring the parameter space remains problematic. Although a very important practical issue, exploration methods have been typically ignored in the literature. A notable exception is in Lin et al. [2010], where the authors admit that their rough estimate of the smooth likelihood is not conducive to parameter estimation.

One of our goals is to ease this computational burden associated with repeatedly obtaining Monte Carlo estimates of the log-likelihood over the parameter space. Much of the previous work has focused on efficiently estimating the likelihood at a fixed $\theta$. An approximate MLE can then be obtained by maximizing this estimated likelihood over $\Theta$. In some cases, the derivatives of the log-likelihood with respect to $\theta$ can be obtained from the simulated values used to produce the estimate of the likelihood as in Stramer and Yan [2007a] and gradient ascent optimization is straightforward. Typically, however, these derivatives must be obtained numerically, which adds a significant computational burden. Although the underlying log-likelihood may be smooth as a function of $\theta$, the Monte Carlo estimates will be subject to variability and thus will be much less amenable to derivative calculation.
Chapter 5: Sequential Kriging-Based Optimization

The references in the previous chapter provide guidance for obtaining estimates at a fixed $\theta$, but exploring the parameter space remains problematic. Although a very important practical issue, exploration methods have been typically ignored in the literature. A notable exception is in Lin et al. [2010], where the authors admit that their rough estimate of the smooth likelihood is not conducive to parameter estimation.

Our approach assumes that the discretized log-likelihood function, $l^{(K)}(\theta)$, is smooth in $\theta$, but that the estimates of this function obtained via the Brownian bridge sampler (our SMC estimate) are subject to Monte Carlo variability. Rather than attempting to maximize the estimated function using a prohibitively large Monte Carlo sample, we propose a sequential optimization method that explicitly models the underlying smooth discretized log-likelihood using a Gaussian process (GP), while treating Monte Carlo variability as measurement error. Using kriging equations, parameter values are added sequentially by maximizing the so-called expected improvement, which balances the uncertainty in estimating the discretized log-likelihood at unexplored parameter values with the desire to find parameter values near the current maximum that have a higher log-likelihood. We call our method sequential kriging-based optimization, or SKBO for short.
5.1 Naive Method

As derivatives of the estimated log-likelihood will typically be unavailable, obtaining an estimated MLE cannot be performed in the usual manner of steepest gradient ascent. A naive solution to this is to choose some fine grid of \( m \) points in each of the \( p \) dimensions of \( \theta \) over the possible parameter values \( \Theta \) and obtain an SMC estimate of the log-likelihood at each of them. The estimated MLE is then obtained by choosing the \( \theta \) value in the grid with the largest estimated log-likelihood.

This method is clearly computationally demanding, as it requires obtaining \( m^p \) estimates of the log-likelihood. Depending on the number of observed data points \( N \) and the specified SMC inputs \( K \) (the number of subintervals between each observation) and \( M \) (the number of Monte Carlo samples used to estimate the transition density), which were discussed in the previous chapter, obtaining each SMC-based estimate of the log-likelihood may not be computationally trivial. Even in cases where the estimates from SMC are relatively easy to obtain, this strategy will quickly suffer the curse of dimensionality as \( p \) grows to even a moderate size. In addition to the computational complexity which initially motivated this work, we also find that the estimated MLE based on the naive approach performs poorly as measured by the bias, standard deviation, and root mean squared error.

We now detail our alternative method of estimating the MLE based on SMC estimates of the log-likelihood function for data from our SDE.
5.2 Gaussian Process Model

5.2.1 Initial Estimates of the Log-Likelihood

Recall from Chapter 2 that we have the process defined in (2.5) as

\[dX(t) = \mu(X(t), \theta) \, dt + \sigma(X(t), \theta) \, dW(t), \quad 0 \leq t \leq T,\]

where \(X(0) = X_0\) is the initial value of the process. The functions \(\mu(\cdot, \cdot)\) and \(\sigma(\cdot, \cdot)\) are assumed to be known up to the \(p\)-dimensional parameter vector \(\theta \in \Theta\), where \(\Theta\) here is assumed to be some compact set in \(\mathbb{R}^p\). This process is observed at time points \(t_i \ (i = 1, \ldots, N)\) where \(0 = t_0 < t_1 < \cdots < t_N\), and we denote these observations by \(X = (X(t_1), \ldots, X(t_N))^\top\). Given \(X\) we start by estimating the discretized log-likelihood

\[l^{(K)}(\theta) = \sum_{i=1}^{N} \log p^{(K)}(X(t_i); X(t_{i-1}), \theta)\]

at a range of parameters values, \(\theta\), that span the space of possible parameter values. For this purpose let \((\theta_1, \ldots, \theta_n)^\top\) denote the initial \(n\) parameter values. There are a variety of choices for \(n\), but the important point is that \(n\) is much smaller than \(m^p\), so that the number initial estimates we obtain is much less computationally demanding than the naive method highlighted in the previous section. At some level, the choice of \(n\) will always be arbitrary as there are a lack of formal results in the literature. A discussion of selecting the number of initial points may be found in Jones et al. [1998]. A common choice is to select \(n = 10p\).

Once the number of initial points, \(n\), has been specified, we use a space-filling design over the parameter space \(\Theta\). A variety of these designs exist in the computer experiments literature based on the competing goals of (1) spacing the points out as
much as possible throughout the parameter space and (2) spacing the points out as 
evenly as possible. We choose the Latin hypercube method, which can be thought 
of as a compromise between the desire to maximize the minimum distance between 
points and the desire to uniformly sample points throughout $\Theta$. A Latin hypercube 
can be thought of as the generalization of the more easily-demonstrated Latin square.

To illustrate a Latin square design with $n$ points, assume a two-dimensional pa-
rameter space with $n \times n$ possible combinations. Then the design is a Latin square if 
and only if there is exactly one sampled point in each row and each column. Table 5.1 
demonstrates two possible Latin squares for a discrete space with $n = 5$.

When the parameter space is continuous, the compact set $\Theta$ (assumed in Chapter 
2) is partitioned similarly to the discrete case, and then points are selected uniformly 
within each square. Table 5.2 demonstrates the corresponding two possible Latin 
squares for a continuous space with $n = 5$. The selected points will be obtained by 
uniformly sampling within each black square. For a Latin hypercube over a continuous 
parameter space, the case in which we are interested, this concept is simply generalized 
to higher dimensions, where there is exactly one sampled point in each hyperplane of 
the parameter space.

To implement this sampling scheme for obtaining $n$ points over a $p$-dimensional 
continuous parameter space $\Theta$, the following steps are taken:

1. For $i = 1, \ldots, p$ do the following:

   (a) Partition the $i^{\text{th}}$ dimension of $\Theta$ into $n$ non-overlapping intervals of equal 
       length.

   (b) Uniformly select a value from each of these $n$ intervals.
Table 5.1: Two possible realizations of a Latin square over discrete space with $n = 5$ points.

Table 5.2: Two possible realizations (corresponding to the discrete realizations in Table 5.1) of a Latin square over continuous space with $n = 5$ points. The sampled points are obtained by uniformly sampling within each of the shaded regions.

2. Randomly combine the $n$ values sampled from the first dimension of $\Theta$ with the $n$ values sampled from the second dimension to obtain $n$ pairs of values.

3. Repeat the previous step for each of the $p$ dimensions to end up with $n p$-tuplets, which constitute the Latin hypercube sample.

Many software packages implement this strategy. We choose to use the function `latin.hypercube` from the `emulator` R package [Hankin, 2005] to obtain a Latin hypercube over $\Theta$. 
Figure 5.1: Realization of Latin hypercube used to obtain 10 initial points for SKBO estimation of MLE for OU process with true parameter $\theta_0 = 2$ and $\theta_1 = -3$.

In the context of using SKBO to estimate the MLE of simulated data from an OU process, one possible set of 10 initial points using a Latin hypercube is displayed in Figure 5.1. It is based on these points that we will initially fit the Gaussian process explained in the following section, which provides the foundation for SKBO.

5.2.2 Gaussian Process Specification

Letting $Y (\theta_i)$ denote the SMC-based estimate of $l^{(K)} (\theta_i)$ at parameter value $\theta_i$ for $i = 1, \ldots, n$ we assume that

$$ Y (\theta_i) = l^{(K)} (\theta_i) + \epsilon (\theta_i) , \quad (5.2) $$

where $\{\epsilon (\theta_i) : i = 1, \ldots, n\}$ is a set of independent $N(0, \sigma^2)$ errors.

Motivated by the belief that the underlying log-likelihood is smooth in $\theta$, we model $\{l^{(K)} (\theta) : \theta \in \Theta\}$ using a Gaussian process (GP) with mean function $\mu_L (\theta; \beta)$
and some valid covariance function \( C_L(\theta, \theta'; \zeta) \), where \( \beta, \zeta \) are unknown parameters. (See Rogers [1985] for a proof of the smoothness of the likelihood function of a one-dimensional SDE satisfying the Lipschitz continuity and linear growth conditions provided in Chapter 2.)

There is an extensive literature on the choice of mean and covariance function for the GP [see Santner et al., 2003, Cressie and Wikle, 2011] – richer choices can more accurately emulate the discretized log-likelihood, at a cost of necessitating larger sample sizes, \( n \), and more computational resources to faithfully model the features of the GP. The mean function may be as simple as the constant mean

\[
\mu_L(\theta; \beta) = \beta
\]

or could be generalized to include covariates. For example, assuming that the mean is linear in the parameter ordinates \( \theta_i, i = 1, \ldots, p \) we get

\[
\mu_L(\theta; \beta) = \beta_0 + \beta_1 \theta_1 + \cdots + \beta_p \theta_p.
\]

Similarly, the covariance function can take on many forms as long as it is an even function and the \( n \times n \) matrix whose \( i, j \) entry is \( C_L(\theta_i, \theta_j; \zeta) \) must be nonnegative definite for all values of \( n \) and choices of \( \theta_i (i = 1, \ldots, n) \). Examples of covariance functions include:

- The Gaussian covariance function, given by

\[
C_L(\theta, \theta'; \zeta) = \zeta_0^2 \exp \left( - \sum_{i=1}^{p} \frac{(\theta_i - \theta_i')^2}{\zeta_i} \right).
\]
The cubic covariance function, given by
\[
C_L(\theta, \theta'; \zeta) = \zeta_0^2 \prod_{i=1}^p \left\{ 2 \left( 1 - \frac{\theta_i - \theta_i'}{\zeta_i} \right)^3 I \left( \frac{\zeta_i}{2} < \theta_i - \theta_i' < \zeta_i \right) + \left[ 1 - 6 \left( \frac{\theta_i - \theta_i'}{\zeta_i} \right)^2 + 6 \left( \frac{\theta_i - \theta_i'}{\zeta_i} \right)^3 \right] I \left( \theta_i - \theta_i' < \frac{\zeta_i}{2} \right) \right\},
\]
where \( I(\cdot) \) is the indicator function.

The Matérn covariance function, given by
\[
C_L(\theta, \theta'; \zeta, \nu) = \zeta_0^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left( \frac{\sum_{i=1}^p (\theta_i - \theta_i')^2}{\sqrt{2\nu} \zeta_1} \right)^\nu K_\nu \left( \sqrt{2\nu} \frac{\sum_{i=1}^p (\theta_i - \theta_i')^2}{\zeta_1} \right),
\]
where \( \Gamma \) is the gamma function and \( K_\nu \) is the modified Bessel function of the second kind of order \( \nu \).

In the examples given above, the covariance functions are all stationary. That is, they depend on the distance between the parameter values \( \theta \) and \( \theta' \), but not on the values themselves. While this is a common assumption, it is not required. The first two covariance functions assume that the underlying GP is smooth; the smoothness of the GP with a Matérn covariance function is determined by the parameter \( \nu \). The larger the value of \( \nu \) the smoother the GP – we obtain the Gaussian covariance as the smoothness parameter \( \nu \to \infty \).

The GP model is popular in both the spatial statistics and computer experiments literature. While the spatial statistics literature is commonly concerned with understanding two and three-dimensional GPs observed with error, the computer experiments literature often deals with situations in much higher dimensions. Typically the output modeled in computer experiments are a smooth deterministic function of the inputs which emulates some real world phenomena (e.g. a nuclear explosion or
the performance of a knee replacement). Obtaining the model output often takes a substantial amount of time, so the computer experiments literature is typically concerned with optimally choosing the inputs to the emulator to determine the “best” inputs to the model as quickly as possible.

Depending on the method used to sequentially select new $\theta$ values at which to obtain a new SMC estimate, it may be possible to obtain multiple estimates of the log-likelihood at the same $\theta$ value. This may also be by design, for example obtaining repeated estimates at a handful of $\theta_i$ values to obtain a pilot estimate of $\sigma^2$. Note that this may be an important practical concern in many situations, as there will be identifiability issues when there is only one estimate $Y(\cdot)$ for each $\theta$ value. That is, without replication, it will be difficult to distinguish between the covariance of the GP $l^{(K)}(\cdot)$ and the variability due to measurement error $\epsilon(\cdot)$.

To allow for the possibility of replication, assume that we have observed $T_i$ SMC estimates at $\theta_i$ and let $T$ be an $n \times n$ diagonal matrix such that $\text{diag}(T) = (T_1, \ldots, T_n)$. Note that when $\theta$ values are sequentially added based on some criterion that is continuous over $\theta$ and the choice is made not use replication a priori, $T$ is equivalent to $I$, the $n \times n$ identity matrix.

Given $Y_n = (Y(\theta_1), \ldots, Y(\theta_n))^\top$, where $Y(\theta_i)$ denotes the average of the estimates over replicates at $\theta_i$, if necessary, we now consider predicting $l^{(K)}(\theta^*)$. Note that $\theta^*$ may or may not be in $(\theta_1, \ldots, \theta_n)^\top$. Given the observed data and the GP parameters $\beta, \zeta$, we obtain the distribution of $l^{(K)}(\theta^*)|Y_n, \beta, \zeta, \sigma^2$ by first considering the joint distribution

$$
\begin{bmatrix}
  l^{(K)}(\theta^*) \\
  Y_n
\end{bmatrix} | \beta, \zeta, \sigma^2 \sim N_{n+1}\left(
\begin{bmatrix}
  \mu_L(\theta^*) \\
  \mu_L(\theta_1, \ldots, \theta_n)
\end{bmatrix},
\begin{bmatrix}
  C_L(\theta^*, \theta^*) & c_L^\top \\
  c_L & \Sigma_L + \sigma^2T^{-1}
\end{bmatrix}
\right).
$$
Standard conditional Gaussian results give best linear unbiased prediction of \( \ell^{(K)}(\theta^*) \) as the kriging mean

\[
\eta_L(\theta^*) = \mu_L(\theta^*) + c_L^T(\Sigma_L + \sigma^2T^{-1})^{-1}[Y_n - \mu_L],
\]

with kriging variance

\[
v_L^2(\theta^*) = C_L(\theta^*,\theta^*) - c_L^T(\Sigma_L + \sigma^2T^{-1})^{-1}c_L.
\]

In the above equations \( \mu_L \) is a mean vector of length \( n \) with \( i \)th element \( \mu_L(\theta_i;\beta) \), \( c_L \) is a covariance vector of length \( n \) with \( i \)th element \( C_L(\theta^*,\theta_i;\zeta) \), and \( \Sigma_L \) is the \( n \times n \) covariance matrix with \((i,j)\) element \( C_L(\theta_i,\theta_j;\zeta) \).

### 5.3 Gaussian Process Parameters

#### 5.3.1 Estimation

We now turn to estimating the parameters \( \beta \) and \( \zeta \) in the mean and covariance functions \( \mu_L(\theta;\beta) \) and \( C_L(\theta,\theta';\zeta) \), respectively. In addition, we wish to estimate the Monte Carlo variability, \( \sigma^2 \). These parameters may be updated in a variety of ways, a few of which are illustrated below.

Conditionally on \( \beta,\zeta,\sigma^2 \), the joint distribution of \( Y_n \) is

\[
Y_n | \beta,\zeta,\sigma^2 \sim N_n \left( \mu_L, \Sigma_L + \sigma^2T^{-1} \right).
\]

Note that \( \text{Var} \ (Y_n | \beta,\zeta,\sigma^2) \) is given by

\[
\Sigma_Y \equiv \Sigma_L + \sigma^2T^{-1},
\]

which can be rewritten as

\[
\Sigma_Y = \sigma^2A,
\]

98
where defining

\[ \Sigma^*_L = \frac{1}{\sigma^2} \Sigma_L, \]

we have

\[ A = T^{-1} + \frac{1}{\sigma^2} \Sigma_L = T^{-1} + \Sigma^*_L. \]

We write \( \Sigma_Y \) in this form so that we can factor the variance of \( Y_n \) into two distinct pieces: the measurement error, \( \sigma^2 \), and the matrix \( A \), which contains sum of the contribution to the variance of \( Y_n \) due to to replication and the covariance matrix of \( l(\theta_1), \ldots, l(\theta_n) \) rescaled by the measurement error. For instance, if \( C_L(\theta_i, \theta_i) = \tau^2 \), then we will reparameterize \( A \) in terms of \( \alpha = \tau^2/\sigma^2 \) so that we can consider \( \sigma^2 \) separately from \( \alpha \) and any remaining covariance parameters \( \zeta \). Assuming that the mean is linear in \( \beta \) (assumed to be a \( q \times 1 \) vector), we can write

\[ \mu_L = H\beta, \]

where the design matrix \( H_{n \times q} = [h(\theta_1), \ldots, h(\theta_n)]^T \) contains the covariates observed at each \( \theta_i \), denoted by \( h(\theta_i) \).

We can then write the likelihood of the data as

\[
L(\beta, \zeta, \sigma^2|Y_n) = \frac{|\Sigma_Y|^{-1/2}}{(2\pi)^{n/2}} \exp \left(-\frac{1}{2} (Y_n - H\beta)^\top \Sigma_Y^{-1} (Y_n - H\beta) \right) = \frac{|A|^{-1/2}}{(2\pi \sigma^2)^{n/2}} \exp \left(-\frac{1}{2\sigma^2} (Y_n - H\beta)^\top A^{-1} (Y_n - H\beta) \right), \tag{5.5}\]

where \(|\cdot|\) is defined to be the determinant of a matrix.

**Maximum Likelihood Estimation**

The maximum likelihood estimates can be found [see, for example Andrianakis and Challenor, 2009] by first obtaining the weighted least squares estimate of \( \beta \) which
corresponds to the maximum likelihood estimate. We wish to maximize (5.5) with respect to $\beta$, which is equivalent to maximizing

$$\log L(\beta, \zeta, \sigma^2 | Y_n) = \log \left( \frac{|A|^{1/2}}{(2\pi\sigma^2)^{n/2}} \right) - \frac{1}{2\sigma^2} (Y_n - H\beta)^\top A^{-1} (Y_n - H\beta)$$

$$\propto - \frac{1}{2\sigma^2} (Y_n - H\beta)^\top A^{-1} (Y_n - H\beta).$$

We then have

$$\frac{d}{d\beta} \log L(\beta, \zeta, \sigma^2 | Y_n) = \frac{d}{d\beta} \left\{ - \frac{1}{2\sigma^2} \left[ (Y_n^\top A^{-1} Y_n) - (Y_n^\top A^{-1} H\beta) 
- ((H\beta)^\top A^{-1} Y_n) + (H\beta)^\top A^{-1} H\beta \right] \right\}$$

$$= - \frac{1}{2\sigma^2} (-2H^\top A^{-1} Y_n + 2HA^{-1} H\beta).$$

Setting this equation equal to zero and solving yields

$$- \frac{1}{2\sigma^2} (-2H^\top A^{-1} Y_n + 2HA^{-1} H\beta) = 0.$$

Thus,

$$HA^{-1} H\beta = H^\top A^{-1} Y_n,$$

and we find that

$$\hat{\beta}_{MLE} = (HA^{-1} H)^{-1} H^\top A^{-1} Y_n.$$

Also note that the second derivative

$$\frac{d^2}{d\beta^2} \log L(\beta, \zeta, \sigma^2 | Y_n) = -\frac{1}{\sigma^2} HA^{-1} H < 0,$$

so that $\hat{\beta}_{MLE}$ is indeed a maximum.
Once \( \hat{\beta}_{\text{MLE}} \) has been obtained, we can substitute this value for \( \beta \) in the likelihood (5.5) to get the profile likelihood

\[
L\left( \hat{\beta}_{\text{MLE}}, \zeta, \sigma^2 \mid \mathbf{Y}_n \right) = \frac{|\mathbf{A}|^{-1/2}}{(2\pi \sigma^2)^{n/2}} \times \exp \left( -\frac{1}{2\sigma^2} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)^\top \mathbf{A}^{-1} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right) \right),
\]

which can then be maximized with respect to \( \sigma^2 \) to find \( \hat{\sigma}^2_{\text{MLE}} \). We have

\[
\frac{d}{d\sigma^2} \log L \left( \hat{\beta}_{\text{MLE}}, \zeta, \sigma^2 \mid \mathbf{Y}_n \right) \propto -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)^\top \mathbf{A}^{-1} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right).
\]

Setting this equation equal to zero and solving yields

\[
\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)^\top \mathbf{A}^{-1} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)
\]

which implies that

\[
\hat{\sigma}^2_{\text{MLE}} = \frac{\left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)^\top \mathbf{A}^{-1} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)}{n}.
\]

Also note that the second derivative evaluated at the MLE is

\[
\frac{d^2}{d(\sigma^2)^2} \log L \left( \beta, \zeta, \sigma^2 \mid \mathbf{Y}_n \right) \bigg|_{\sigma^2 = \hat{\sigma}^2_{\text{MLE}}} = -\frac{n}{2 \left( \hat{\sigma}^2_{\text{MLE}} \right)^2} - \frac{1}{3 \left( \hat{\sigma}^2_{\text{MLE}} \right)^3} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)^\top \mathbf{A}^{-1} \left( \mathbf{Y}_n - \mathbf{H}\hat{\beta}_{\text{MLE}} \right)
\]

\[
= -\frac{n}{2 \left( \hat{\sigma}^2_{\text{MLE}} \right)^2} - \frac{n}{\left( \hat{\sigma}^2_{\text{MLE}} \right)^2}
\]

\[
= -\frac{n}{2 \left( \hat{\sigma}^2_{\text{MLE}} \right)^2} < 0,
\]

so that \( \hat{\sigma}^2_{\text{MLE}} \) is truly a maximum.
The final step is to find the MLEs for the remaining GP parameters $\zeta$. These are given by

$$\hat{\zeta}_{\text{MLE}} = \arg \max_{\zeta} \left[ L \left( \hat{\beta}_{\text{MLE}}, \zeta, \hat{\sigma}^2_{\text{MLE}} \mid Y_n \right) \right],$$

and these MLEs will typically be difficult to obtain depending on the functional form of $C_L (\theta, \theta'; \zeta)$, as they do not generally exist in closed-form. (Typically numerical methods are required to obtain $\hat{\zeta}_{\text{MLE}}$.)

### Restricted Maximum Likelihood Estimation

An additional problem with using the MLEs to estimate the GP parameters arises due to the problem pointed out in Stein [1999] that “$H\hat{\beta}_{\text{MLE}}$ will always be closer to $Y_n$ than to $H\beta$; so the MLE of $\zeta$ will underestimate the variation in the process at least relative to what we would get if we knew $\beta$”. The restricted MLEs [Patterson and Thompson, 1971], which address this issue and properly account for the uncertainty in $\beta$, maximize the restricted likelihood given by

$$L \left( Y_n \mid \zeta, \sigma^2 \right) = \int_{\beta} L \left( Y_n \mid \beta, \zeta, \sigma^2 \right) d\beta \propto \frac{|A|^{-1/2}|H^\dagger A^{-1} H|^{-1/2}}{(2\pi\sigma^2)^{(n-q)/2}} \times \exp \left( -\frac{1}{2\sigma^2} \left( Y_n - H\hat{\beta}_{\text{MLE}} \right)^\dagger A^{-1} \left( Y_n - H\hat{\beta}_{\text{MLE}} \right) \right).$$

The restricted likelihood may also be obtained in alternatively as the likelihood of the parameters for a linear contrast of $Y_n$ which is independent of $\beta$. Similar calculations to those in the case of the unrestricted MLEs yield

$$\hat{\sigma}^2_{\text{RMLE}} = \frac{(Y_n - H\hat{\beta}_{\text{MLE}})^\dagger A^{-1} (Y_n - H\hat{\beta}_{\text{MLE}})}{n - q},$$

102
\[
\widehat{\zeta}_{\text{RMLE}} = \arg \max_{\zeta} \left[ L \left( Y_n | \zeta, \hat{\sigma}^2_{\text{RMLE}} \right) \right].
\]

Again, \( \widehat{\zeta}_{\text{RMLE}} \) will typically not exist in closed-form and requires numerical solution.

**Bayesian Approach**

Bayesian methods provide the advantage of allowing us to account for all sources of uncertainty, which includes the uncertainty about the estimated parameters. In addition, the Bayesian approach provides the ability to directly incorporate information gleaned from pilot estimates of the measurement error, \( \sigma^2 \), into the prior distribution \([\sigma^2]\) rather than simply plugging in the pilot estimate. We adopt this Bayesian viewpoint and specify a prior distribution \([\beta, \zeta, \sigma^2]\) for the GP parameters. Throughout the remainder of this dissertation we use the square bracket notation \([\cdot]\) to denote “distribution of”. Using Bayes’ rule,

\[
[\beta, \zeta, \sigma^2 | Y_n] \propto [\beta, \zeta, \sigma^2][Y_n | \beta, \zeta, \sigma^2]
= [\beta, \zeta, \sigma^2]|\Sigma_Y|^{-1/2} \exp \left\{ -\frac{1}{2} (Y_n - \mu_L)^\top \Sigma_Y^{-1} (Y_n - \mu_L) \right\}.
\]

The posterior distribution (5.6) can be explored and summarized in several ways. In our examples below, we take our estimates \((\hat{\beta}_n, \hat{\zeta}_n, \hat{\sigma}^2_n)\) to be the posterior mode of (5.6). Empirically, this method led to the best results as far as accurately representing the log-likelihood over \( \Theta \) and guiding the sequential search strategy that we introduce later in this chapter.

We can also investigate the posterior distribution of the covariance parameters given the data (integrating out \( \beta \)). This distribution allows us to perform Bayesian estimation of the covariance parameters in conjunction with *any* method of estimating
\( \beta \) (e.g. MLE or OLS). We saw that obtaining estimates of \( \beta \) was relatively easy compared to obtaining estimates of the covariance parameters of the GP, so in practical settings the user may often be free to estimate the mean of the GP using any method above and then may need to resort to Bayesian methods for the covariance parameters out of convenience. Let \( \omega = (\zeta, \sigma^2)^T \) be the covariance parameter vector (the GP parameters except for \( \beta \)). For simplicity assume that the GP has the constant mean function \( \mu_L = \beta 1_n \). The GP with constant mean has been shown to perform well in terms of interpolation, in our setting predicting within \( \Theta \) [Santner et al., 2003]. If the space \( \Theta \) is unknown or misspecified initially, then including the elements of \( \theta \) as covariates can be useful for extrapolating, or indicating a more appropriate space to explore. We retain the simpler case.

We write

\[
[\omega|Y_n] = [\omega, Y_n] | Y_n[\omega] \propto [Y_n|\omega][\omega]
\]

\[
= \int \left[ Y_n, \beta | \omega \right][\omega]d\beta = \int \left[ Y_n|\beta, \omega \right][\omega][\beta]d\beta.
\]

That is, we can rewrite the posterior distribution of the covariance parameters \( \omega \) given the observed log-likelihood values \( Y_n \) as the product of the conditional distribution of \( Y_n \) given \( \omega \) and the prior distributions of \( \omega \) and the mean parameter \( \beta \), with \( \beta \) integrated out.

Assuming the uninformative prior for \( \beta : [\beta] \propto 1 \) and assuming that \( \beta \) is independent of \( \omega \), we have

\[
[\omega|Y_n] \propto [\omega] \int \left[ Y_n|\beta, \omega \right]d\beta.
\]
Since we are assuming that

$$[Y_n|\beta, \omega] = N_n(\beta 1_n, \Sigma_Y)$$

we have the posterior distribution of the covariance parameters given the current set of estimated log-likelihoods is

$$\omega[Y_n] \propto \omega |\Sigma_Y|^{-1/2} \int \exp \left\{ -\frac{1}{2} (Y_n - \beta 1_n)^T \Sigma_Y^{-1} (Y_n - \beta 1_n) \right\} d\beta$$

$$= \omega |\Sigma_Y|^{-1/2} \int \exp \left\{ -\frac{1}{2} (Y_n^T \Sigma_Y^{-1} Y_n - \beta 1_n^T \Sigma_Y^{-1} \beta 1_n + \beta 2^T 1_n \Sigma_Y^{-1} 1_n) \right\} d\beta$$

$$= \omega |\Sigma_Y|^{-1/2} \int \exp \left\{ -\frac{1}{2} (Y_n^T \Sigma_Y^{-1} Y_n) \right\}$$

$$\times \int \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[ \left( \beta - \frac{1_n^T \Sigma_Y^{-1} Y_n}{1_n^T \Sigma_Y^{-1} 1_n} \right)^2 - \left( \frac{1_n^T \Sigma_Y^{-1} Y_n}{1_n^T \Sigma_Y^{-1} 1_n} \right)^2 \right] \right\} d\beta$$

$$= \omega |\Sigma_Y|^{-1/2} \int \exp \left\{ -\frac{1}{2} \left( Y_n^T \Sigma_Y^{-1} Y_n - \frac{1_n^T \Sigma_Y^{-1} Y_n}{1_n^T \Sigma_Y^{-1} 1_n} \right)^2 \right\} d\beta$$

$$= \omega \frac{|\Sigma_Y|^{-1/2}}{\sqrt{1_n^T \Sigma_Y^{-1} 1_n}} \exp \left\{ -\frac{1}{2} \left( Y_n^T \Sigma_Y^{-1} Y_n - \frac{1_n^T \Sigma_Y^{-1} Y_n}{1_n^T \Sigma_Y^{-1} 1_n} \right)^2 \right\}$$

(5.7)

where

$$\int \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \left( \beta - \frac{1_n^T \Sigma_Y^{-1} Y_n}{1_n^T \Sigma_Y^{-1} 1_n} \right)^2 \right\} d\beta = \sqrt{1_n^T \Sigma_Y^{-1} 1_n}$$
results from recognizing that we have a univariate Gaussian kernel with mean and variance
\[
\frac{1}{n} \Sigma_Y^{-1} Y_n \quad \text{and} \quad \frac{1}{n} \Sigma_Y^{-1} 1_n.
\]

Maximization of the Posterior Distribution

From a practical perspective, maximizing the posterior distribution in (5.6) or (5.7) may prove difficult. Note that the \( n \times n \) matrix \( \Sigma_Y \) must be recalculated for different \( \omega \) values. In the examples considered later in this dissertation, we found that maximizing the log-posterior distribution in terms of \( \log(\omega) \) could typically be done using a linear optimization procedure such as the Nelder-Mead method [Nelder and Mead, 1965]. Depending on the size of the number of \( \theta \) values, \( n \), this may not be a trivial computation.

5.4 Expected Improvement

Let \( \tilde{\eta}_L = \max_{i=1, \ldots, n} \eta_L (\theta_i) \) denote the maximum value of the kriging mean over the explored \( \theta \) values. Then, the improvement at \( \theta^* \) is [Jones et al., 1998]
\[
I(\theta^*) = \max \left\{ 0, l^{(K)} (\theta^*) - \tilde{\eta}_L \right\}, \quad (5.8)
\]
but since \( l^{(K)} (\theta^*) \) is unknown we replace it by the expected improvement at parameter value \( \theta^* \), which can be shown to be equal to [Jones et al., 1998]
\[
E(I(\theta^*) | Y_n) = \left[ \eta_L (\theta^*) - \tilde{\eta}_L \right] \Phi \left( \frac{\eta_L (\theta^*) - \tilde{\eta}_L}{v_L (\theta^*)} \right) + v_L (\theta^*) \phi \left( \frac{\eta_L (\theta^*) - \tilde{\eta}_L}{v_L (\theta^*)} \right), \quad (5.9)
\]
where recall that \( \phi(\cdot) \) is the standard normal probability density function and \( \Phi(\cdot) \) is the standard normal cumulative distribution function. As explained above, the
expected improvement balances the need to maximize the discretized log-likelihood (the first term) and the uncertainty in estimating the log-likelihood (the second term).

Our sequential optimization scheme adds the parameter value \( \theta^* \) that maximizes the expected improvement (5.9); details are given at the end of this section. We then estimate the discretized log-likelihood using SMC at that new parameter value, yielding \( Y(\theta^*) \) and update the data \( Y_{n+1} = (Y_n^T, Y(\theta^*))^T \) and the GP parameter estimates \( \hat{\beta}_{n+1}, \hat{\zeta}_{n+1} \). After updating the kriging mean (5.3) and kriging variance (5.4) using \( Y_{n+1}, \hat{\beta}_{n+1}, \hat{\zeta}_{n+1} \), we search for another parameter value that maximizes the expected improvement. We continue observing new estimated log-likelihood values and optimizing (5.9) to find more parameter values, until some stopping criteria is met.

It is then straightforward to obtain the estimated MLE, \( \hat{\theta} = \arg \max_{i=1,...,n} \eta_L(\theta_i) \), where \( n \) is the total number of points observed after the final iteration in this procedure.

**Maximizing the expected improvement.** Maximizing (5.9) may be done using either an optimization procedure (e.g. Nelder-Mead) or over a grid. If an optimization approach is used, to aid the exploration of \( E(I(\cdot)|Y_n) \) over \( \theta \), the derivative of (5.9) with respect to \( \theta^* \) can be shown to be

\[
\frac{d}{d\theta^*}E(I(\theta^*)|Y_n) = \frac{d\eta_L(\theta^*)}{d\theta^*} \Phi \left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right) + \frac{dv_L(\theta^*)}{d\theta^*} \phi \left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right).
\]

(5.10)

The derivatives \( d\eta_L(\theta^*)/d\theta^* \) and \( dv_L(\theta^*)/d\theta^* \) will depend on the mean and covariance functions of the chosen GP and may be calculated accordingly. The grid approach may be simpler to implement, but it should be noted that this will become computationally intensive as the dimension of \( \theta \) grows. This will be similar to the problem with estimating \( l^{(K)}(\cdot) \) over a fine grid, but it will occur at a slower rate because the computations involved in calculating expected improvement are simpler.
than those used to estimate the discretized log-likelihood. If a grid approach is used, adjustments should also be made to allow for replicates at a given $\theta_i$.

**Choosing a stopping rule.** We choose to stop adding points once we have observed no change to $\hat{\theta}$ for five consecutive added points. As noted in Williams et al. [2000], expected improvement is not monotonically decreasing as points are added, so choosing a stopping rule may prove difficult. Possible alternative choices for stopping criteria may be based on a fixed number of points $n$ or small changes in $\eta_L(\cdot)$. Alternatively, consecutively observing the maximum expected improvement over $\Theta$ below some threshold is a viable strategy as well.

A run of SKBO is displayed in Figure 5.2 for illustrative purposes. This plot shows the true log-likelihood as contour lines and the initial points as squares for the same $N = 1000$ observations from the OU process used to illustrate the Latin hypercube. The SKBO algorithm was stopped when the change to $\hat{\theta}$ is less than 0.01 in each dimension for five consecutive iterations, which led to the log-likelihood being estimated at 36 points in addition to the initial estimates selected based on the Latin hypercube. For a given iteration, the most recently added point is marked by a filled red circle, while the previously added points are displayed as gray circles. The true MLE calculated based on these 1000 observations is displayed using solid lines, and the estimated MLE based on the estimates at the sampled points at the current iteration is displayed using dashed lines. Selected iterations are used to demonstrate the evolution of the estimated MLE and the sampled points via SKBO through time.
Figure 5.2: Initial points (squares), previously added points (gray circles), point added via SKBO on current iteration (red filled circle), estimated MLE on current iteration (dashed line), and true MLE (solid line) for $N = 1000$ observations $\Delta = 0.1$ apart in time simulated from an OU model with true parameter values $\theta_0 = 2$ and $\theta_1 = -3$. The SMC-based estimates were obtained using $K = 5$ fill-in points and $M = 25$ Monte Carlo samples for each pair of adjacent observations. The SKBO algorithm was stopped when the change to $\hat{\theta}$ was less than 0.01 for five consecutive iterations, which was after 36 points had been added in this illustrative run.
5.4.1 Approximate Confidence Intervals

In the usual setting, we can obtain an approximate $(1 - \alpha)\%$ joint confidence region for $\theta$ based on the fact that as the number of observed data values $N$ grows large enough, for the true parameter value $\theta_0$ and the MLE $\hat{\theta}$,

$$\left[ \frac{l(\hat{\theta})}{l(\theta_0)} \right]^2 = 2 \left[ l(\hat{\theta}) - l(\theta_0) \right] \sim \chi^2_p,$$

where $p$ is the dimension of $\theta$ and $\chi^2_p$ represents the chi-squared distribution with $p$ degrees of freedom. This is then used to construct the confidence region for $\theta$ given by

$$\left\{ \theta : 2 \left( l(\hat{\theta}) - l(\theta) \right) \leq \chi^2_{1 - \alpha, p} \right\}, \quad (5.11)$$

where $\chi^2_{1 - \alpha, p}$ is the $1-\alpha$ quantile of a chi-square distribution with $p$ degrees of freedom.

In our setting, we do not have access to $l(\theta)$, but if $K$ is chosen large enough so that $l(\theta) \approx l^{(K)}(\theta)$, we can write

$$\left\{ \theta : 2 \left( l^{(K)}(\hat{\theta}) - l^{(K)}(\theta) \right) \leq \chi^2_{1 - \alpha, p} \right\}.$$

Again, since $l^{(K)}(\theta)$ is unknown, we replace it by its conditional expectation to obtain an approximate $(1 - \alpha)\%$ joint confidence region for $\theta$ directly from the kriging mean:

$$\left\{ \theta : 2 \left( \eta_L(\hat{\theta}) - \eta_L(\theta) \right) \leq \chi^2_{1 - \alpha, p} \right\}. \quad (5.12)$$

A visual demonstration of (5.11) and (5.12) is given in Figure 5.3 for $N = 1000$ observations from an OU process (2.19) with $\theta_0 = 2$ and $\theta_1 = -3$. We note that the two confidence regions both follow the general shape of the true log-likelihood (displayed as gray contour lines) and that there is a large overlap between the two.
The difference between the two arises due to the fact that the kriging mean estimates the log-likelihood to be flatter than it truly is, and thus the confidence region based on the kriging mean includes more $\theta$ values.

Given the GP parameters $\eta_L(\cdot)$ will be trivial to compute relative to obtaining an estimate $Y(\cdot)$. However as the dimension of $\theta$, $p$, grows obtaining the region in (5.12) may become difficult; in that case we can use a Rao-based confidence region of the form

$$\left\{ \theta : \left( \hat{\theta} - \theta \right)^\top \hat{I}(\hat{\theta})^{-1} \left( \hat{\theta} - \theta \right) \leq \chi^2_{1 - \alpha, p} \right\},$$  

where we can estimate the Fisher information, $\hat{I}(\hat{\theta})$ using the second derivative of $\eta_L(\cdot)$ with respect to $\theta$ evaluated at $\hat{\theta}$. While this will depend on the specific GP parameters chosen, we demonstrate the calculations for a specific functional form of $\eta_L(\cdot)$ later in this chapter.

### 5.5 Maximizing Expected Improvement

#### 5.5.1 General Case

Recall that

$$E(I(\theta^*)|Y_n) = [\eta_L(\theta^*) - \tilde{\eta}_L] \Phi\left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right) + v_L(\theta^*) \phi\left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right),$$

and that we wish to find $\frac{d}{d\theta^*}E(I(\theta^*)|Y_n)$.

Applying the product rule to each of the two terms above yields

$$\frac{d}{d\theta^*}E(I(\theta^*)|Y_n) = \frac{d\eta_L(\theta^*)}{d\theta^*} \Phi\left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right) + \left[ \eta_L(\theta^*) - \tilde{\eta}_L \right] \frac{d\Phi\left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right)}{d\theta^*}$$

$$+ \frac{dv_L(\theta^*)}{d\theta^*} \phi\left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right) + v_L(\theta^*) \frac{d\phi\left( \frac{\eta_L(\theta^*) - \tilde{\eta}_L}{v_L(\theta^*)} \right)}{d\theta^*}.$$
Figure 5.3: Approximate confidence regions based on the likelihood ratio test for $N = 1000$ observations from an OU process with $\theta_0 = 2$ and $\theta_1 = -3$ constructed using the true log-likelihood (blue) and kriging mean (red).
Using the chain rule and the fact that for a normal density,
\[ \frac{d\phi(x)}{dx} = -x\phi(x), \]
we have
\[ \frac{d}{d\theta^*} E(I(\theta^*)|Y_n) = \frac{d\eta_L(\theta^*)}{d\theta^*} \Phi\left( \frac{\eta_L(\theta^*) - \bar{\eta}_L}{\nu_L(\theta^*)} \right) \]
\[ + \left[ \eta_L(\theta^*) - \bar{\eta}_L \right] \phi\left( \frac{\eta_L(\theta^*) - \bar{\eta}_L}{\nu_L(\theta^*)} \right) \frac{d\eta_L(\theta^*) - \bar{\eta}_L}{d\theta^*} \]
\[ + \frac{dv_L(\theta^*)}{d\theta^*} \phi\left( \frac{\eta_L(\theta^*) - \bar{\eta}_L}{\nu_L(\theta^*)} \right) \]
\[ - \frac{\nu_L(\theta^*)}{v_L(\theta^*)} \phi\left( \frac{\eta_L(\theta^*) - \bar{\eta}_L}{\nu_L(\theta^*)} \right) \frac{d\eta_L(\theta^*) - \bar{\eta}_L}{d\theta^*}. \]

Noting that the second and fourth terms cancel, this yields
\[ \frac{d}{d\theta^*} E(I(\theta^*)|Y_n) = \frac{d\eta_L(\theta^*)}{d\theta^*} \Phi\left( \frac{\eta_L(\theta^*) - \bar{\eta}_L}{\nu_L(\theta^*)} \right) \]
\[ + \frac{dv_L(\theta^*)}{d\theta^*} \phi\left( \frac{\eta_L(\theta^*) - \bar{\eta}_L}{\nu_L(\theta^*)} \right). \quad (5.14) \]

### 5.5.2 Specific Mean and Covariance Functions

We now examine \( d\eta_L(\theta^*)/d\theta^* \) and \( dv_L(\theta^*)/d\theta^* \) for a given choice of mean and covariance functions. We consider
\[ \mu_L(\theta) = \beta \text{ and } C_L(\theta_1, \theta_2) = \tau^2 \exp\{(\theta_1 - \theta_2)^\top(\theta_1 - \theta_2)/\eta\}. \]

We first compute the \( p \times n \) matrix
\[ \frac{d_{\theta^*}^\top}{d\theta^*} = \frac{d}{d\theta^*} \left[ C_L(\theta^*, \theta_1), \ldots, C_L(\theta^*, \theta_n) \right]_{1 \times n} \]
\[ = -\frac{2}{\eta} \left[ (\theta^* - \theta_1)^\top C_L(\theta^*, \theta_1), \ldots, (\theta^* - \theta_n)^\top C_L(\theta^*, \theta_n) \right]_{p \times n}. \quad (5.15) \]

Now we obtain the \( p \times 1 \) vectors
\[ \frac{d\eta_L(\theta^*)}{d\theta^*} = \frac{d\mu_L(\theta^*)}{d\theta^*} + \frac{d}{d\theta^*} (c_L^\top (\Sigma_L + \sigma^2 I_n)^{-1} [Y_n - \mu_L]) \]
\[ = \frac{dc_L^\top}{d\theta^*} (\Sigma_L + \sigma^2 I_n)^{-1} [Y_n - \mu_L]. \quad (5.16) \]
and

\[
\frac{dv_L(\theta^*)}{d\theta^*} = \frac{dc^T_L dv_L(\theta^*)}{d\theta^*} \frac{dc^T_L}{dc^T_L} \\
= \frac{dc^T_L}{d\theta^*} \left( \frac{d}{dc^T_L} \sqrt{\tau^2 - c^T_L (\Sigma_L + \sigma^2 I_n)^{-1} c_L} \right) \\
= \frac{dc^T_L}{d\theta^*} \left( (\tau^2 - c^T_L (\Sigma_L + \sigma^2 I_n)^{-1} c_L)^{-1/2} c^T_L (\Sigma_L + \sigma^2 I_n)^{-1} \right)^T.
\]

(5.17)

We now have the necessary pieces to calculate \(d/d\theta^* E(I(\theta^*)|Y_n)\), which will aid in efficiently maximizing expected improvement over \(\theta\). This will determine the next value of \(\theta\) at which an estimate of the log-likelihood is to be obtained via the chosen SMC method. In conjunction with updating the GP parameters using one of the methods laid out in this chapter (which may be done less frequently as \(n\) increases and the GP parameters become more stable), these steps provide the foundation of SKBO at each iteration. In the following chapter, we investigate the accuracy of the estimated MLE based on simulation studies and demonstrate the flexibility of SKBO using historical stock market data.
Chapter 6: Simulation Results and Application to Modeling Stock Price

6.1 Simulated Results

We now use data simulated from two models to evaluate the performance of the sequential kriging-based optimization (SKBO) compared to naive space-filling designs in terms of accuracy and speed. We consider a “practical” naive space-filling design which estimates $l^{(K)}(\cdot)$ at $25p$ points across $\Theta$. We choose to stop the sequential search when the estimate of $\hat{\theta}$ changes by less than .01 in each direction for five consecutive iterations, or when we have sampled $25p$ points, whichever occurs first, to provide a fair comparison to the naive method.

A common rule-of-thumb in the computer experiments literature is to use $n = 10p$ initial $Y(\cdot)$ values based on a space-filling design. To investigate the role that the initial number of points plays, we perform SKBO based on $n = 5p$ and $n = 10p$ initial $Y(\cdot)$ values. We also attempted to investigate the alternative strategy of using a large ($M = 25000$) Monte Carlo sample to guide a steepest-ascent search, but even at this value of $M$, in which the estimated $Y(\cdot)$ values took 1000 times as long to obtain, the log-likelihood is not smooth enough for the optimization to converge.
As the choices of $K$ and $M$ will affect the shape of $l^K(\cdot)$ and the Monte Carlo variability which we estimate using $\sigma^2$, we consider two combinations of these values for each model. This practical choice is necessary for any SMC-based estimation, regardless of the strategy used to search $\Theta$. As the difference between $l(\cdot)$ and $l^K(\cdot)$ is of order $1/K$ [Bally et al., 1995], $K$ should be selected to be large enough to reduce this bias to some acceptable level, while keeping computational cost in mind. Once $K$ has been selected, we choose $M = K^2$, which is the most efficient use of computational resources for the modified Brownian bridge sampler [Stramer and Yan, 2007b]. To avoid the difficulties with SMC that can be caused for moderate or large $N$, which were illustrated in Chapter 4, we run pilot studies with $M = 2K^2$ and $M = 4K^2$ to ensure that the estimates do not contain any downward bias. Although Pedersen’s sampler encountered difficulties with this bias at $N = 1000$, based on pilot runs for each of the models we consider, and the independent investigation visualized in Figure 4.12, we do not find any evidence that the modified Brownian bridge sampler of Durham and Gallant [2002] suffers from this bias. We note that when $K$ and $M$ are specified to be large, it is even more important to carefully explore the parameter space, as each $Y(\cdot)$ value is more difficult to obtain.

In the following simulations we choose

$$
\mu_L(\theta) = \beta \quad \text{and} \quad R(\theta, \theta') = \exp \left( -||\theta - \theta'||^2 / \eta \right).
$$

We also assume

$$
[\beta, \zeta] \propto \frac{\eta}{(\sigma^2 + \tau^2)}.
$$
where \( \beta \in \mathbb{R} \) and \( \sigma^2, \tau^2, \eta > 0 \). We find that even when these choices do not reflect reality, the sequential kriging-based optimization performs well. We used R [R Core Team, 2013] for the implementation of SKBO and sped up the matrix calculations involved in calculating the expected improvement using C++ through the R package RcppArmadillo [Eddelbuettel and Sanderson, 2014]. We also used C to quickly obtain the modified Brownian bridge sample based on an adaptation of code in the R package sde [Iacus, 2009b], and the latin.hypercube function of the R package emulator [Hankin, 2005] to obtain the initial grid over \( \Theta \).

### 6.1.1 Ornstein-Uhlenbeck Process

We first consider the Ornstein-Uhlenbeck (OU) model [Uhlenbeck and Ornstein, 1930] given by

\[
\begin{align*}
    dX_t &= (\theta_0 + \theta_1 X_t) \, dt + dW_t, \quad 0 \leq t \leq T, \\
    X_0 &= x_0
\end{align*}
\]

where \( X_0 = x_0 \) is the initial value of the process, \( \theta_0 \in \mathbb{R}, \theta_1 < 0 \), and \( W_t \) is a standard Brownian motion. Without loss of generality, we have assumed unit diffusion in (6.1). If this is not the case, (6.1) may be transformed into a process with unit diffusion via the Lamperti transform [see Iacus, 2009a, for example]. This process has a transition density that is available in closed-form, so we can use the true log-likelihood and MLE to evaluate the performance of SKBO and the naive method.

We obtain our simulated data by first simulating \( X_0 \) from the stationary distribution of (6.1), given by

\[
p(X_0, \theta) = \phi \left( X_0; \theta_0/\theta_1, 1/\sqrt{2\theta_1} \right).
\]

We then sequentially simulate \( X_{t_1}, \ldots, X_{t_N} \) from the conditional distribution of (6.1), which is given by

\[
p(X_{\Delta}|X_0, \theta) = \phi \left( X_{\Delta}; X_0 e^{-\theta_1(1-\Delta)} + \frac{\theta_0}{\theta_1} \left[ 1 - e^{-\theta_1(1-\Delta)} \right], \frac{1}{2\theta_1} \left[ 1 - e^{-2\theta_1(1-\Delta)} \right] \right),
\]

(6.2)
Figure 6.1: (a) A contour plot of the discretized log-likelihood for the OU process given by (6.1) with $\theta_0 = 2$ and $\theta_1 = -3$. The solid horizontal and vertical lines denote the exact MLEs of $\theta_0$ and $\theta_1$ respectively, and the dashed lines denote the SKBO-based estimate. For the SKBO method, the squares indicate the initial parameter values, and the circles denote the values added sequentially. (b) is a zoomed in version of (a).

where $N$ is chosen to be 1000 and $t_i - t_{i-1} = 0.1$, for $i = 1, \ldots, N$. We use $\theta_0 = 2$ and $\theta_1 = -3$ for this analysis. After repeating the analysis for various combinations of $\theta_0$ and $\theta_1$, we found the results presented below to be insensitive to the choice of these parameters.

We first investigate how well the SKBO method finds the maximum for one simulated realization from (6.1). Panels (a) and (b) of Figure 6.1 display, as a contour plot, the discretized log-likelihood for the realization of the OU process (Panel (b) is a zoomed in version of (a)). Looking at the figure, we can see that the likelihood is concentrated along a line of positive slope in the $(\theta_0, \theta_1)$ space. In Figure 6.1, the squares denote the initial sample of 10 points generated using a Latin hypercube and
the circles denote the $\theta$ values added using SKBO method, maximizing the expected improvement as each new parameter value is added. The majority of the circles are concentrated along a ridge of high log-likelihood in the parameter space, which indicates that the method is able to isolate an estimate of the MLE. A few circles further away from the ridge have been added to decrease the uncertainty in estimating the discretized log-likelihood using the GP. On the figure the horizontal and vertical lines denote the exact MLEs (solid line) and SBO-based estimates (dashed line) of $\theta_0$ and $\theta_1$, respectively. The estimates are close for this realization.

Now we evaluate the general performance of SKBO-based estimation of $\theta$ relative to the exact MLE and naive space-filling methods. Table 6.1 compares the performance of each method as we vary the accuracy of the SMC approximation for the naive and SKBO methods (controlled by $K$ and $M$ – we compare $K = 5$ with the $K = 10$ case) and as we change the initial number of points sampled in the parameter space. For this simple process with $p = 2$ parameters we are able to add a more computationally intensive design which samples 2500 points, a number that could prove to be computationally impractical for other processes.

In Table 6.1 a number of different criteria are compared. To compare the quality of the estimates we summarize the bias, standard deviation (SD), and root mean square error (RMSE) for each method. The bootstrap standard error for the bias, SD, and RMSE is bounded above by 0.02. In addition to recording the average time to find the approximate MLE, and the number of points used initially for the SMC-based methods, we present the average number of points added for the SKBO method.

Comparing to the gold standard of the exact MLE, as expected, the approximate SMC-based methods (naive and SKBO methods) have a larger bias, SD, and RMSE.
Table 6.1: A comparison of the true MLE for 1000 simulations of an OU process given by (6.1) with the MLEs obtained using two SMC-based methods: the “Naive” space-filling method and the “SKBO” method. For each SMC-based method we vary $K$ and $M$ to control the accuracy of the SMC approximation. We also vary the number of points sampled in the parameter space. The bootstrap standard error for the bias, SD, and RMSE is bounded above by 0.02.

<table>
<thead>
<tr>
<th></th>
<th>MLE</th>
<th>$K = 5, M = 25$</th>
<th></th>
<th>$K = 10, M = 100$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Naive</td>
<td>SKBO</td>
<td>Naive</td>
<td>SKBO</td>
</tr>
<tr>
<td>Initial pts</td>
<td>–</td>
<td>50</td>
<td>2500</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Avg added</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Bias</td>
<td>0.03</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>0.23</td>
<td>0.56</td>
<td>0.42</td>
<td>0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.22</td>
<td>0.56</td>
<td>0.43</td>
<td>0.33</td>
<td>0.29</td>
</tr>
<tr>
<td>Bias</td>
<td>0.06</td>
<td>0.12</td>
<td>0.13</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>0.28</td>
<td>0.66</td>
<td>0.55</td>
<td>0.42</td>
<td>0.36</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.29</td>
<td>0.67</td>
<td>0.57</td>
<td>0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>Coverage</td>
<td>98.1%</td>
<td>–</td>
<td>–</td>
<td>75.6%</td>
<td>84.5%</td>
</tr>
<tr>
<td>Avg Time</td>
<td>–</td>
<td>20.5</td>
<td>1030.7</td>
<td>20.8</td>
<td>18.6</td>
</tr>
</tbody>
</table>
Accounting for the uncertainty of these measures, the SKBO method outperforms the naive space-filling methods, regardless of the choice of the number of initial points. Compared to the SKBO method the 2500-point naive method still has disappointing performance. Also, the SKBO method takes much less time and uses significantly fewer points (at least 32% fewer points on average) to get a better estimate than the naive space-filling method. In terms of obtaining a better estimate of the MLE using the SKBO method, we do better for a larger initial number of points, and we improve as the accuracy of the SMC approximation improves. Note that as we increase the approximation accuracy we need slightly less parameter values in total.

We also compared the coverage of 95% confidence regions for $\theta$ using the exact MLE with our SKBO-based method. The results are presented in the last line of Table 6.1. With 1000 replicates, testing that the coverage is equal to 95%, a $\alpha = 0.05$ rejection region for coverages is below 96.3% and above 96.4%. Thus the coverage for the exact MLE of 98.1% is too high. The SKBO-based coverages are too low, but the coverages increase to 88.6% as we increase the accuracy of the SMC approximation and we increase the initial number of points used. The accuracy of the SKBO-based estimate of the discretized log-likelihood will depend on the degree of difference between the specified GP model and the true log-likelihood. By focusing on the log-likelihood near the maximum, the shape of the confidence regions based on SKBO will be heavily influenced by the shape of the log-likelihood near the estimated peak. To the degree that the behavior of the log-likelihood near the maximum is not reflective of its behavior elsewhere, the confidence regions will be anti-conservative. Increasing the number of initial points alleviates this problem.
6.1.2 Generalized CIR Model

We now consider the generalized Cox-Ingersoll-Ross (GCIR) model, introduced in Chan et al. [1992], and analyzed in Roberts and Stramer [2001]. The process is defined by

\[ dX_t = (\theta_0 + \theta_1 X_t) \, dt + \gamma X_t^\psi \, dW_t, \quad 0 \leq t \leq T, \quad (6.3) \]

where \( X_0 = x_0 \) is the initial value of the process, \( \theta_0, \theta_1 \in \mathbb{R}, \gamma > 0, \psi \in [0, 1] \), and \( W_t \) is a standard Brownian motion. We note that (6.3) does not have a closed-form likelihood, except for when \( \psi = 0 \) (the OU process) or \( \psi = 0.5 \) (the Cox-Ingersoll-Ross model). To improve our exploration of the parameter space, we let

\[ \theta_2 = \log(\gamma) \text{ and} \]
\[ \theta_3 = \log\left(\frac{\psi}{1 - \psi}\right), \]

and optimize over the real-valued parameters \( \theta = (\theta_0, \theta_1, \theta_2, \theta_3) \).

To simulate from the GCIR process we first simulate 100000 data points at time increments of 0.001 from (6.3) based on the Euler approximation with \( \theta_0 = 0.5, \theta_1 = -0.25, \gamma = 1, \) and \( \psi = 0.75 \). Next, subsampling every 100 time points, we obtain a realization of length \( N = 1000 \) and sampling interval \( \Delta = 0.1 \).

Table 6.2 compares the naive and SKBO methods for approximating the MLEs for the GCIR process. The format of the table is similar to that of Table 6.1, with the exception that we are not able to calculate an exact MLE. The table illustrates that the naive method and SKBO perform similarly with respect to estimating \( \theta_2 \) and \( \theta_3 \) when \( K = 5 \), but the SKBO method appears to slightly outperform the naive method when \( K = 10 \) by these measures. For estimating \( \theta_0 \) and \( \theta_1 \), similar to the
Table 6.2: A comparison of the estimated MLEs from 1000 simulations of a GCIR process given by (6.3) with the MLEs obtained using two SMC-based methods: the “Naive” space-filling method and the “SKBO” method. For each SMC-based method we vary $K$ and $M$ to control the accuracy of the SMC approximation. We also vary the number of points sampled in the parameter space. The bootstrap standard error for the bias, SD, and RMSE is bounded above by 0.03.

<table>
<thead>
<tr>
<th></th>
<th>$K = 5, M = 25$</th>
<th>$K = 10, M = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive</td>
<td>SKBO</td>
</tr>
<tr>
<td>Initial pts</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Avg added</td>
<td>–</td>
<td>60.3</td>
</tr>
<tr>
<td>$\theta_0$</td>
<td>Bias</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.79</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>Bias</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>1.11</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>Bias</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.27</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>Bias</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.85</td>
</tr>
<tr>
<td>Avg Time</td>
<td>176.4</td>
<td>183.7</td>
</tr>
</tbody>
</table>
OU case, SKBO noticeably outperforms the naive method, regardless of the accuracy of the SMC approximation. It is difficult to distinguish between the performance of the SKBO method on the basis of the number of initial points used. With a smaller initial design we evaluate fewer points on average, but the timing, bias, SD, and RMSE values are harder to discriminate. This suggests that the performance of the SKBO method can be improved by increasing the initial number of points up to some fixed number, but that there is relatively small benefit in increasing the initial number of points past this point. Overall we stress that this number (roughly 20 in both the OU and GCIR cases) is very small relative to the number of points required for the naive space-filling method.

6.2 Stock Data

In this section we apply the methodology described above to the stock price of three companies: Apple, Inc. (AAPL), Hewlett-Packard Co. (HPQ), and Yahoo! Inc. (YHOO). We consider the daily adjusted closing price of each stock from the ten-year period of May 17th, 2004 through May 16th, 2004 for a total of 2518 observations each. The data, shown graphically in Figure 6.2 were obtained from the Historical Prices feature of Yahoo! Finance (http://finance.yahoo.com – accessed May 17th, 2014).

An overwhelmingly popular model for stock price is given by the geometric Brownian motion (GBM) process defined in (2.18). Stock prices were modeled by (2.18) in the famous Black-Scholes model [Black and Scholes, 1973] for option pricing and (2.18) is referred as “the model for stock prices” in Hull [2012], a popular introductory finance text. To initially measure the fit of this model we can use the fact that if $X_t$
Figure 6.2: Daily adjusted closing price for Apple, Inc., Hewlett-Packard Co., and Yahoo! Inc.
follows (2.18) then \( \log(X_t) - \log(X_{t-1}) \) follows a normal distribution. In Figure 6.3 we have the normal Q-Q plot of the log return, which plots the theoretical quantiles of a normal distribution vs. the corresponding quantiles of the observed data. When the points on this plot do not roughly follow a straight line, this indicates a departure from normality. We see that for all three of the stocks in Figure 6.3, the distribution of log return appears to have heavier tails than what would be expected if the data were normally distributed.

We choose \( \Delta = 1/252 \) as there are roughly 252 trading days per year. As (2.18) has a known transition density, we can compute the MLEs and the corresponding log-likelihood for each stock. These values are displayed in the left of Table 6.3.

As a possible alternative model, we consider the generalized GBM given by the solution to

\[
dX_t = \theta_0 X_t \, dt + \gamma X_t^\psi \, dW_t, \quad 0 \leq t \leq T,
\]

where \( X_0 = x_0 \) is the initial value of the process, \( \theta_0 \in \mathbb{R}, \gamma > 0, \psi \in [0, 1] \), and \( \{W_t\} \) is a standard Brownian motion. That is, we investigate whether the model for these three stock prices can be improved by allowing \( \psi \) to vary. As (6.4) does not have a known transition density, we estimate the MLEs for each stock using SKBO with \( K = 10 \) and \( M = 100 \). Once these estimated MLEs have been obtained, we estimate the log-likelihood using the modified Brownian bridge sampler with \( K = 20 \) and \( M = 400 \). This highlights the important point that while the estimated MLE may be obtained for fairly small values of \( K \) and \( M \), estimating the log-likelihood accurately will typically require larger values. Also, while \( \eta_L(\cdot) \) is very useful for guiding the search of \( \Theta \), we do not use it to estimate \( l(\hat{\theta}) \) as this would essentially be estimating the maximum using a mean.
Figure 6.3: Normal Q-Q plot of log return for Apple, Inc., Hewlett-Packard Co., and Yahoo! Inc.

<table>
<thead>
<tr>
<th>Stock</th>
<th>$\hat{\theta}$</th>
<th>$\hat{\gamma}$</th>
<th>Log-lik.</th>
<th>AIC</th>
<th>$\hat{\theta}$</th>
<th>$\hat{\gamma}$</th>
<th>$\hat{\psi}$</th>
<th>Est. Log-lik.</th>
<th>AIC</th>
<th>LRT Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>0.45</td>
<td>0.37</td>
<td>-6797</td>
<td>13598</td>
<td>0.34</td>
<td>1.09</td>
<td>0.79</td>
<td>-6740</td>
<td>13486</td>
<td>114</td>
</tr>
<tr>
<td>HPQ</td>
<td>0.12</td>
<td>0.33</td>
<td>-2341</td>
<td>4686</td>
<td>0.30</td>
<td>1.27</td>
<td>0.61</td>
<td>-2331</td>
<td>4668</td>
<td>20</td>
</tr>
<tr>
<td>YHOO</td>
<td>0.10</td>
<td>0.40</td>
<td>-2149</td>
<td>4302</td>
<td>0.37</td>
<td>1.03</td>
<td>0.71</td>
<td>-2135</td>
<td>4276</td>
<td>28</td>
</tr>
</tbody>
</table>
As we would expect since the models are nested, the log-likelihood of the data modeled by the generalized GBM is higher than that modeled by the GBM for each of the three stocks. Evaluating the two models based on AIC (two times the number of parameters minus the log-likelihood), we see that the generalized GBM model outperforms the usual GBM model for each of these three stocks. Using the fact that the models are nested, we can also compute the likelihood ratio statistic to test the null hypothesis that the GBM model fits sufficiently well versus the alternative hypothesis that the generalized GBM is required. From Table 6.3, we see that the observed likelihood-ratio test (LRT) statistics are 114, 20, and 28 for AAPL, HPQ, and YHOO, respectively. We obtain the corresponding p-values by comparing the observed statistics to an equally-weighted mixture distribution of a $\chi^2$ distribution with one degree of freedom and a $\chi^2$ distribution with two degrees of freedom. (Note that this differs from the usual asymptotic distribution of the LRT statistic due to the fact that the alternative hypothesis is on the boundary of the null hypothesis. Details may be found in Self and Liang [1987]) As the largest of the three p-values is smaller than $10^{-5}$, we conclude that the geometric Brownian motion model does not sufficiently fit these three datasets and that the parameter $\psi$ should be included in the model.

The alternative model to the usual GBM is only one of many possible extensions. The drift and diffusion terms can be generalized in many different ways, leading to many candidate models. The fit of competing candidate models may be compared, as above, using the estimates of the log-likelihood obtained via SMC, evaluated at the estimated MLEs. The significant contribution of this work is to help determine these points at which the log-likelihood should be estimated in lieu of the naive grid-based
approach and gradient-based approaches, which prove computationally prohibitive in most practical settings.
Chapter 7: Conclusions, Discussion, and Future Work

7.1 Conclusion and Discussions

In this dissertation, we began by briefly discussing the difficulties in performing likelihood inference for observations modeled by SDEs. Many of these difficulties are caused by the fact that, although the models exist in continuous time, the data are necessarily observed discretely in time in practice. The Markov property of these models allows for the likelihood to be written as the product of the individual transition densities for each pair of sequential observations, so that estimated transition densities can be used to construct an estimate of the likelihood. We discussed four classes of estimation methods for the transition density: (1) approximations derived by numerically solving the Kolmogorov forward equation, (2) closed-form Hermite expansions of the transition density, (3) sequential Monte Carlo (SMC) and (4) methods based on the exact simulation of diffusions. After reviewing the existing literature, which compares these methods based on their performance in estimating transition densities, we provided additional comparisons based on simulation studies.

We then considered the performance of the methods in estimating the overall log-likelihood, an important practical problem that has received less attention in the past. We found that as the number of observations grows, and data are thus
observed further in the tails of the conditional distribution given the previous point, each of the methods begins to encounter difficulties. For the Hermite approximation [Aït-Sahalia, 2002a], especially when the data are observed at less frequent intervals, the estimated likelihood can be negative, and thus the estimated log-likelihood is undefined. In Figure 4.3 and Table 4.2, we demonstrated that this problem happens more frequently as the number of observations increases. For $N = 10000$ data values observed $\Delta = 0.1$ apart in time from the “best case” OU model (2.19), the estimates of the log-likelihood were undefined over 80% of the time on average.

While remedying the problems of the other methods is fairly involved, the bias observed in the SMC estimates can be decreased by simply increasing $M$, the number of Monte Carlo samples. This problem is most apparent in Pedersen’s method of estimating the log-likelihood as demonstrated in Figure 4.4 and is caused by the skew of the distribution from which the estimates of the log-likelihood are generated. By tying the simulated variates which are used to calculate the estimated log-likelihood to the end observation, Durham’s and Gallant’s modified Brownian bridge sampler significantly reduces the skew of this distribution. Depending on the number of observations, $N$, and the sophistication of the SDE, the number of Monte Carlo samples, $M$, may still need to be increased for to avoid bias in the modified Brownian bridge sampler, but to a much smaller extent than the adjustment required for Pedersen’s sampler. This was demonstrated visually in Figure 4.12, which compared the shape of the distribution of the estimates based on each of these schemes. In practical situations, the true log-likelihood is unavailable as a tool to check for the downward bias which can appear in the estimated log-likelihoods obtained via SMC when $M$ is too
small. When this is the case, multiple $M$ values should be used to check for convergence to ensure that the bias has been sufficiently reduced. That is, if for example, doubling $M$ changes the variability of the estimates over $\theta$, but does not noticeably shift the values of these estimates, this can indicate that the original $M$ is sufficiently large.

Next, we introduced a sequential, kriging-based optimization strategy which provides a derivative-free method for approximating the MLE in SDE models, in cases where the likelihood function cannot be evaluated exactly, but can be estimated with some degree of statistical accuracy. This work is primarily motivated by the case where the data are discrete-time observations of an SDE; however, what we suggest can be extended to other settings without difficulty. We judge the performance of our approach on two fronts: (1) statistical accuracy, as measured by the bias, SD, and RMSE of the estimated MLE, and (2) computational efficiency. Our findings show that the proposed method outperforms space-filling competitors, even in the cases where these methods use significantly more likelihood evaluations.

### 7.2 Future Work

A significant component of the SKBO method is the assumption that the unobserved likelihood is a realization of a certain GP, specified via its mean and covariance structure. The validity of this assumption is difficult to establish. However, in the cases where we do have access to the exact likelihood function we do not find any evidence to disprove it. On the other hand, the Gaussian assumption offers the significant advantage of yielding a closed form expression for the expected improvement.
In addition, the GP-method allows us to provide measure of uncertainty for our estimates (the MLE and the underlying estimates of the likelihood function). For the examples discussed above, we use simple and popular parametric functions to describe this underlying GP. We show that even a naive formulation leads to impressive results. We do however acknowledge that the parameterization of the GP process may be crucial in certain applications. Although we found the simple functions

\[ \mu_L(\theta) = \beta \] and

\[ R(\theta, \theta') = \exp \left( -\|\theta - \theta'\|^2 / \eta \right) . \]

to be sufficient for our purposes, a natural extension is to consider more sophisticated functions governing the GP. The constant mean generally appears to perform well when the true value of \( \theta \) lies within the specified parameter space \( \Theta \), i.e. interpolation. In practice, however, the specification of \( \Theta \) may often be difficult. When there is significant concern about misspecification of \( \Theta \), covariates can be included in the mean function to indicate areas which should be added to \( \Theta \), i.e. for extrapolation. For the covariance function, it may be helpful to relax the assumption that the covariance is \textit{isotropic}. That is, the covariance function we have used implies that the covariance between two points depends only on the overall distance between the points, not on the distance in each dimension of \( \theta \). For sophisticated models where the dimensions of \( \theta \) have varying importance in determining the shape of the log-likelihood (for example, the GCIR model of Chapter 6), it seems reasonable that the performance of SKBO would be improved by allowing the covariance to reflect this. Additionally, for examples with a more involved log-likelihood function, a non-stationary GP process may be necessary to improve the estimation of the MLE. With more sophisticated GP processes, we would need to sample more parameter values.
to learn about the unknown log-likelihood function and its inherent uncertainty. We take a Bayesian viewpoint to learn and update the GP parameters, thus naturally incorporating uncertainty and allowing us to take advantage of prior knowledge.

From a computational perspective, we identify two bottlenecks: (1) maximizing the expected improvement and (2) updating the structure of the underlying GP given that a new parameter has been added to the procedure. For a specific application one can use state of the art software to reduce the computational overhead in each case. In the applications presented in this paper we implement the SKBO method using multi-core R routines. This is done in order to allow for a fair comparison to other approaches. Our method lends itself naturally to parallel computing, given that we make heavy use of Monte Carlo and importance sampling techniques. Thus, we expect that a more sophisticated approach which uses several Central Processing Units (CPUs) or even a Graphical Processing Unit (GPU) may offer significant improvements.

Modern statistical applications involve observational models with ever increasing complexity. In many cases, an exact evaluation of the likelihood function is impossible and thus one is forced to use an array of approximations. Our proposed method offers several advantages, making it a valuable addition to the MLE toolbox. We are currently investigating its application to the analysis of other statistical models.

Another useful extension would be to adjust $M$ for each pair of sequential observations. As seen in Chapter 3, the bias which can show up when too few Monte Carlo samples are used to obtain estimates of the log-likelihood via SMC are typically caused by a handful of “extreme” observations. Although increasing $M$ for each transition density provides a simple remedy to this problem, it will often be the
case that a lot of this extra computation is unnecessary. Rather than using the same $M$ to estimate each transition density, it may be worthwhile to consider varying the number of Monte Carlo samples based on some estimate of how extreme a given pair of observations are.

Additional future work includes extending SKBO to the Bayesian setting. Given a prior distribution on $\theta$, $\pi(\theta)$, we are often interested in $\pi(\theta|X)$, the posterior distribution of $\theta$ given the observations $X$ from the SDE (2.5), which can be written as

$$
\pi(\theta|X) \propto \pi(\theta)L(\theta|X),
$$

where $L(\theta|X)$ is the likelihood as defined in (2.17). In order to find the posterior mode of this distribution, we can consider the log-posterior distribution given by

$$
\log (\pi(\theta|X)) \propto \log (\pi(\theta)) + l(\theta|X),
$$

where $l(\theta|X)$ is the log of the likelihood. Replacing the true log-likelihood with the discretized log-likelihood $l^{(K)}$ defined in (5.1), we can proceed in a similar fashion to Chapter 5 to estimate the maximum of the discretized posterior, denoted by $\log (\pi^{(K)}(\theta|X))$. One possibility for doing this is to model the SMC-based estimate $Y(\theta_i)$ as

$$
Y(\theta_i) = \log (\pi^{(K)}(\theta|X)) - \log (\pi(\theta)) + \epsilon(\theta_i),
$$

where $\{\epsilon(\theta_i) : i = 1, \ldots, n\}$ is again a set of independent $N(0, \sigma^2)$ errors. The constant variance assumption could be relaxed, as well. That is, for $i = 1, \ldots, n$, we could model each of the independent $\epsilon(\theta_i)$ as a $N(0, \sigma_i^2)$ error.

In the frequentist methods developed in Chapter 5 we were concerned with finding the maximum of a Gaussian Process observed with noise. In the Bayesian context,
we now have the additional piece contributed by the deterministic log-prior. Further investigation is needed to determine how best to specify the prior distribution, but one possibility is to specify the log-prior as a given smooth GP so that $Y(\theta_i)$ is the sum of three components: (1) an unobserved GP which we are interested in maximizing (The new GP of interest will have mean function given by $\mu_L(\theta; \beta) + \log (\pi(\theta))$), (2) a known GP, and (3) random noise. This approach has the advantage of preserving the Gaussianity of $Y_n$, an important fact we used in the frequentist setting to develop SKBO. An alternative way to approach this problem would be to combine the second and third components to obtain a new set of errors $\{\epsilon^*(\theta_i) : i = 1, \ldots, n\}$ which are independent errors normally distributed, again with variance $\sigma^2$, but now with mean given by $-\log (\pi(\theta_i))$ rather than 0.
Bibliography


