Importance sampling techniques for estimation of diffusion models

Omiros Papaspiliopoulos[*], and Gareth O. Roberts [†]

May 22, 2009

1 Overview of the chapter

This article develops a class of Monte Carlo (MC) methods for simulating conditioned diffusion sample paths, with special emphasis on importance sampling schemes. We restrict attention to a particular type of conditioned diffusions, the so-called diffusion bridge processes. The diffusion bridge is the process obtained by conditioning a diffusion to start and finish at specific values at two consecutive times $t_0 < t_1$.

Diffusion bridge simulation is a highly non-trivial problem. At an even more elementary level unconditional simulation of diffusions, that is without fixing the value of the process at t_1, is difficult. This is a simulation from the transition distribution of the diffusion which is typically intractable. This intractability stems from the implicit specification of the diffusion as a solution of a stochastic differential equation (SDE). Although the unconditional simulation can be carried out by various approximate schemes based on discretizations of the SDE, it is not feasible to devise similar schemes for diffusion bridges in general. This has motivated active research in the last 15 years or so for the development of MC methodology for diffusion bridges.

The research in this direction has been fuelled by the fundamental role that diffusion bridge simulation plays in the statistical inference for diffusion processes. Any statistical analysis which requires the transition density of the process is halted whenever the latter is not explicitly available, which is typically the case. Hence it is challenging to fit diffusion models employed in applications to the incomplete data typically available. An interesting possibility is to approximate the intractable transition density using an appropriate MC scheme and carry

[*] Department of Economics, Universitat Pompeu Fabra, Spain, email: omiros.papaspiliopoulos@upf.edu

[†] Department of Statistics, University of Warwick, UK, email: Gareth.O.Roberts@warwick.ac.uk

out the analysis using the approximation. It is of course desirable that the MC scheme is such that the approximation error in the analysis decreases to 0 as the MC effort increases. It turns out that basically all such MC schemes require diffusion bridge simulation.

We have been vague about the nature of "statistical analysis" mentioned above, since a range of statistical problems can be tackled using the diffusion bridge simulation methodology we develop here: parameter estimation of discretely observed diffusions, on-line inference for diffusions observed with error, off-line posterior estimation of partially observed diffusions etc. Additionally a range of computational statistics tools can be combined with the simulation methodology to give answers to the aforementioned problems: the EM algorithm, simulated likelihood, Sequential Monte Carlo, Markov chain Monte Carlo etc. Given the wide range of applications where diffusions are employed, it is not surprising that important methodological contributions in bridge simulation are published in scientific journals in statistics, applied probability, econometrics, computational physics, signal processing etc. Naturally, there is a certain lack of communication across disciplines and one of the aims of this chapter is to unify some fundamental techniques.

In this article we concentrate on two specific methodological components of the wide research agenda described above. Firstly, we derive importance sampling schemes for diffusion bridge simulation. We refer to diffusion bridge simulation as an *imputation* problem, since we wish to recover an unobserved path given its end points. Secondly, we demonstrate how the samples can provide estimators of the diffusion transition density. Such estimators can be directly used in a simulated likelihood framework to yield approximations to the maximum likelihood estimator for the parameters of a discretely observed diffusion. We refer to estimation of the transition density as an *estimation* problem.

A fundamental complication in this context is that the diffusion bridge is an infinite dimensional random variable. One strategy to tackle this issue is to first approximate the stochastic process with a finite-dimensional vector, a so-called skeleton of the bridge obtained at a collection of n intermediate time points in $[t_0, t_1]$. This step adds a further approximation error in the analysis, which to be eliminated n has to be chosen large enough. Subsequently, one has to devise a MC sampling scheme for the corresponding n-dimensional distribution. Let us for convenience call this paradigm the projection-simulation strategy. The problem with this approach is that typically reducing the approximation bias (increasing n) leads to an increase of the MC variance.

An alternative strategy is to design an appropriate MC scheme which operates on the infinite dimensional space, hence in principle it returns diffusion bridges. In our specific framework, we are interested in importance sampling for diffusion bridges. There is a certain mathematical hurdle in this direction since it requires changes of measure in infinite dimensional spaces. For practical implementation we might have to approximate the output of the simulation algorithm using a skeleton of the bridge based on n intermediate points. Again, the approximation bias is eliminated as $n \to \infty$. Let us refer to this paradigm as the simulation-projection strategy.

There are two main advantages of this strategy over the projection-simulation. Firstly, it often results in a much better bias/variance tradeoff. Roughly speaking, the fact that by construction a valid MC scheme exists in the limit $n \to \infty$ avoids the curse of dimensionality from which they often suffer the projection-simulation strategies. Secondly, in some contexts the approximation step is unnecessary and unbiased MC methods can be devised. The retrospective simulation technique is instrumental in this context, and from an operational point of view the output of the algorithm consists of a skeleton of the diffusion bridge unvailed at a collection of random times.

The organisation of the chapter is as follows. In Section 2 essential introductory material is presented on diffusions and Monte Carlo methods. Sections 3 and 3 consider the fundamental problem of Importance Sampling and related methods for diffusion bridges. Finally in Section 5, the use of all the methodology of previous sections in conjunction with exact and unbiased methods for diffusion simulation are briefly described.

2 Background

2.1 Diffusion processes

Diffusion processes are extensively used for modelling continuous-time phenomena in many scientific areas; an indicative list includes economics (Merton, 1971; Bergstrom, 1990), finance (Cox et al., 1985; Sundaresan, 2000; Chan et al., 1992) biology (McAdams and Arkin, 1997; Wilkinson, 2006), genetics (Kimura and Ohta, 1971; Tan, 2002), chemistry and physics (van Kampen, 1981), dynamical systems (Arnold, 1998; Givon et al., 2004) and engineering (Bar-Shalom et al., 2002). Their appeal lies in the fact that the model is built by specifying the instantaneous mean and variance of the process through a stochastic differential equation (SDE). Specifically, a d-dimensional diffusion process $V \in R^d$ is a strong Markov process defined as the solution of an SDE of the type:

$$\mathrm{d}V_s = b(s, V_s)\,\mathrm{d}s + \sigma(s, V_s)\,\mathrm{d}B_s, \quad s \in [0, T]\,, V_0 = v_0; \tag{1}$$

B is an m-dimensional standard Brownian motion, $b(\cdot\,,\cdot) : R_+ \times R^d \to R^d$ is called the *drift*, $\sigma(\cdot\,,\cdot) : R_+ \times R^d \to R^{d \times m}$ is called the *diffusion coefficient*. We will treat the initial point v_0 as fixed by the design, although it is straightforward to model it with a distribution on R^d. In applications the drift and the diffusion matrix are only known up to some parameters, which have to be estimated. It is convenient to introduce also $\Gamma = \sigma\sigma^*$. We assume that all coefficients are sufficiently regular so that (1) has a unique weak non-explosive solution, and (crucially) that the Cameron-Martin-Girsanov theorm holds. Details of these conditions can be found for example in Rogers and Williams (2000).

The popularity of SDEs in time-series modelling is due to various reasons: they provide a flexible framework for modelling both stationary processes with quite general invariant distributions and non-stationary processes; in many applications they arise as limits of discrete-time and/or discrete-space Markov processes; they are a natural stochastic counterpart to deterministic modelling using Ordinary Differential Equations (ODEs); the Markov property is particularly convenient from a computational perspective allowing fast statistical inference for long time-series; smooth processes can also be modelled by allowing the drift and diffusion coefficient to depend on the past of the process.

The SDE (1) describes the *microscopic* behaviour of the process, i.e. its dynamics in infinitesimal time increments. One the other hand, the *exact macroscopic* dynamics of the diffusion process are governed by its transition density:

$$p_{s,t}(v,w) = \mathrm{P}\left[\, V_t \in \mathrm{d}w \;\mid\; V_s = v \,\right] / \,\mathrm{d}w, \quad t > s,\ w, v \in R^d \,. \tag{2}$$

There are very few examples of SDEs with tractable transition densities. One generic class of such processes is the so-called linear SDEs, where the drift is linear in the state variable and the diffusion matrix is constant with respect to the state variable, see the corresponding subsection of this section. This class incorporates the Ornstein-Uhlenbeck process, a special case of which is a model due to Vasicek (1977) for the term structure of interest rates. Also in the context of interest rate modelling Cox et al. (1985) proposed a non-linear SDE, which however has a known transition density.

Although typically intractable, the transition density has various representations which suggest different approaches for its approximation. We could identify two main representations. First, it is given as a solution of the Fokker-Planck partial differential equation (PDE) with appropriate initial and boundary conditions. There are various methods to solve the PDE numerically, see for example Hurn and Lindsay (2007) for a recent article which investigates this possibility in the context of estimation of diffusions. Second, it can be expressed in various ways as an expectation, and these expressions lend themselves to Monte Carlo approximation. It is this second approach which is pursued in this article and linked with diffusion bridge simulation.

Numerical approximation

The core of Monte Carlo methodology for SDEs is the simulation of a skeleton of the process $\{V_{t_0}, V_{t_1}, \ldots, V_{t_n}\}$. In fact, there are two types of simulations which can be considered for SDEs. Simulating the strong solution essentially corresponds to jointly constructing V, the solution of (1), and B, the Brownian motion which drives it. On the other hand, simulating the weak solution only asks for simulating V according to the probability law implied by (1).

Note that due to the strong Markov property, *exact simulation* of a skeleton entails sequential simulation from the transition density (2), which however is typically intractable. Exact simulation of strong solutions is clearly an even harder problem. Nevertheless, a vast collection of *approximate simulation* schemes are available based on discretizations of the SDE (1). The simplest approximation scheme is the Euler-Murayama approximation (Maruyama, 1955):

$$V_{t+\Delta t} \approx V_t + b(t, V_t)\Delta t + \sigma(t, V_t)\sqrt{\Delta t} \cdot (B_{t+\Delta t} - B_t)\,. \tag{3}$$

In this form the discretisation tries to approximate the diffusion in a *strong* sense. A weak Euler method is not constrained to have Gausian innovations, but it will have innovations with the correct mean and variance (up to a certain order of Δt). Let V_T^Δ denote an approximate solution based on a Δ-discretization of $[0, T]$. We say that a strong approximation scheme is of order γ if

$$\mathrm{E}[|V_T - V_T^\Delta|] \le K\Delta^\gamma$$

and correspondingly a weak approximation scheme is of order γ if

$$|\mathrm{E}[g(V_T)] - \mathrm{E}[g(V_T^\Delta)]| \le K\Delta^\gamma$$

for suitable test functions g. Under suitable regularity conditions on the coefficients of the SDE, a strong Euler scheme is of order $1/2$ in the strong sense, whereas in general Euler schemes are of order 1 in the weak sense. Many higher order schemes exist. Some are based on the Itô-Taylor expansion, and there are also implicit and split-step methods (which are particularly important in the construction of MCMC methods using diffusion dynamics). For a detailed exposition of numerical approximation of SDEs we refer to Kloeden and Platen (1995).

Diffusion bridges

We now consider the dynamics of the process V not only conditioned on its initial point, $V_0 = u$, but also on its ending point, $V_T = v$. The conditioned process, which we will also denote by V (the distinction from the unconditioned process will be clear from the context), is still Markov, and the theory of *h-transforms* (see for example Rogers and Williams, 2000, Chapter IV.39) allows us to derive its SDE:

$$\begin{aligned} \mathrm{d}V_s &= \tilde{b}(s, V_s)\,\mathrm{d}s + \sigma(s, V_s)\,\mathrm{d}B_s, \quad s \in [0, T]\,, V_0 = u; \\ \tilde{b}(s, x) &= b(s, x) \,+\, [\sigma\sigma^*](s, x)\,\nabla_x \log p_{s,T}(x, v) \end{aligned} \tag{4}$$

where σ^* denotes the matrix transpose. There are three main remarks on this representation. First, note that the local characteristics of the unconditioned and conditioned processes are the same, in the sense that they share the same diffusion coefficient. Second, the drift of the conditioned process includes an extra term which forces the process to hit v at time T. Third, although insightful for the bridge dynamics, (4) is typically intractable since the drift is expressed in terms of the transition density.

Therefore, the diffusion bridge solves an SDE whose drift is intractable, hence even approximate simulation using the schemes described above is infeasible. Application of that technology would require first the approximation of the transition density, and consequently a discretization of (4). This is clearly impractical, and it is difficult to quantify the overall error of the approach. Instead, we will consider alternative Monte Carlo schemes for simulating diffusion bridges.

Data and likelihood

In some contexts we can observe directly a *path* of the modelled process $V = (V_s,\, s \in [0,T])$. More realistically we might be able to observe a *skeleton* of the process $\{V_{t_0}, V_{t_1}, \ldots, V_{t_n}\}$, but where the frequency of the data can be chosen arbitrarily high. A typical example is molecular dynamics modelling, where the data is simulated according to a complex deterministic model (Stuart and Pokern, 2009, see for example). A rich mathematical framework is available for statistical analyses in this high frequency regime, see for example Prakasa Rao (1999). Two main components of this theory is the quadratic variation identity and the Cameron-Martin-Girsanov change of measure. According to the former, the local characteristics of the SDE can be completely identified given an observed path. In particular, for any $t \in [0,T]$,

$$\lim_{\Delta \to 0} \sum_{t_j \leq t} (V_{t_{j+1}} - V_{t_j})(V_{t_{j+1}} - V_{t_j})^* = \int_0^t [\sigma\sigma^*](s, V_s)\mathrm{d}s \tag{5}$$

in probability for any partition $0 = t_0 \leq t_1 \leq \cdots \leq t_n = t$, whose mesh is Δ. This implies that from high frequency data we can consistently estimate the diffusion coefficient. A further implication is that the probability laws which correspond to SDEs with different diffusion coefficient are mutually singular. On the contrary, under weak conditions the laws which correspond to SDEs with the same diffusion coefficient but different drifts are equivalent and a simple expression for the Radon-Nikodym derivative is available. This is the context of the Cameron-Martin-Girsanov theorem for Itô processes, see for example Theorem 8.6.6 of Øksendal (1998).

In the context of (1) consider functionals h and α of the dimensions of b and assume that u solves the equation:

$$\sigma(s,x)h(s,x) = b(s,x) - \alpha(s,x)\,.$$

Additionally, let $\mathbb{P}_b$ and $\mathbb{P}_\alpha$ be the probability laws implied by the (1) with drift b and α respectively. Then, under certain conditions $\mathbb{P}_b$ and $\mathbb{P}_\alpha$ are equivalent with density (*continuous time likelihood*) on $\mathcal{F}_t = \sigma(V_s, s \leq t)$, $t \leq T$, given by

$$\left.\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\alpha}\right|_t = \exp\left\{\int_0^t h(s,V_s)^*\mathrm{d}B_s - \frac{1}{2}\int_0^t [h^*h](s,V_s)\mathrm{d}s\right\}\,. \tag{6}$$

In this expression, B is the $\mathbb{P}_\alpha$ Brownian motion, and although this is the usual probabilistic statement of the Cameron-Martin-Girsanov theorem, it is not a natural expression to be used in statistical inference, and alternatives are necessary. For example, note that when σ can be inverted, the expression can be considerably simplified. Recall that $\Gamma = \sigma\sigma^*$, then the density becomes

$$\exp\left\{\int_0^t [(b-\alpha)^*\Gamma^{-1}](s, V_s)\mathrm{d}V_s - \frac{1}{2}\int_0^t [(b-\alpha)^*\Gamma^{-1}(b+\alpha)](s, V_s)\mathrm{d}s\right\} . \quad (7)$$

Appendix contains a simple presentation of change of measure for Gaussian multivariate distributions, which might be useful for the intuition behind the Girsanov theorem. For statistical inference about the drift (1) the Girsanov theorem is used with $\alpha = 0$. Any unknown parameters in the drift can be estimated by using (6) as a likelihood function. In practice, the integrals in the density are approximated by sums, leading to an error which can be controlled provided the data are available at arbitrarily high frequency.

Nevertheless, in the majority of applications V can only be partially observed. The simplest case is that of a *discretely observed diffusion*, where we observe a skeleton of the process $\{V_{t_0}, V_{t_1}, \ldots, V_{t_n}\}$, but without any control on the frequency of the data. As a result, the approach described above is not feasible since it might lead to large biases (see for example Dacunha-Castelle and Florens-Zmirou, 1986). From a different persective, we deal with data from a Markov process, hence the joint Lebesgue density of a sample (*discrete time likelihood*) is simply given by the product of the transition densities

$$\prod_{i=0}^{n-1} p_{t_i,t_{i+1}}(V_{t_i}, V_{t_{i+1}}) . \quad (8)$$

Unknown parameters in the drift and diffusion coefficient can be estimated working with this discrete-time likelihood. Theoretical properties of such estimators are now well known in particular under ergodicity assumptions, see for example Kessler (1997); Gobet (2002); Kutoyants (2004). Unfortunately the discrete-time likelihood is not practically useful in all those cases where the transition density of the diffusion is analytically unavailable.

More complicated data structures are very common. In many applications V consists of many components which might not be synchroneously observed, or there the observation might be subject to measurement error, or there might be components completely latent. However, likelihood estimation of discretely observed diffusions will serve as a motivating problem throughout the subsequent methodological sections.

Linear SDEs

A great part of the diffusion bridge simulation methodology is based on the tractability of linear SDEs and uses this class as a building block. Hence, it is useful to include a short description of this class. This is a large family of SDEs characterised by state-independent diffusion coefficient and drift which is linear in the state variable. In the most general form we have

$$\mathrm{d}V_s = (D(s)\, V_s + G(s))\,\mathrm{d}s + E(s)\,\mathrm{d}B_s\,; \tag{9}$$

hence $b(s,x) = D(s)x + G(s)$, and $\sigma(s,x) = E(s)$, where D, G, E are matrix-valued functions of appropriate dimensions, which are allowed to depend only on time. Linear SDEs constitute one of the few families of equations which can be solved analytically. We define $P(s)$ to be the solution of the linear ODE

$$\frac{\mathrm{d}P}{\mathrm{d}s} = D(s)P(s)\,, \quad P(0) = I\,. \tag{10}$$

Then the SDE is solved by

$$V_t = P(t)\int_0^t P(s)^{-1}(G(s)\mathrm{d}s + E(s)\mathrm{d}B_s) + P(t)v_0\,. \tag{11}$$

It follows that V is a *Gaussian process* with mean $m_t := \mathrm{E}(V_t)$ and covariance matrix $C_t := \mathrm{Cov}(V_t, V_t)$, which solve the systems of ODEs

$$\begin{aligned}\frac{\mathrm{d}m_t}{\mathrm{d}t} &= D(t)m_t + G(t)\,, \quad m_0 = v_0\\ \frac{\mathrm{d}C_t}{\mathrm{d}t} &= D(t)C_t + C_tD(t)^* + \Gamma(t)\,, \quad C_0 = 0\,.\end{aligned}$$

where $\Gamma(s) = E(s)E(s)^*$. In various contexts these ODEs can be solved analytically.

The transition density is Gaussian with mean and variance derived from the previous expressions. A consequence of this is a further appealing feature of linear SDEs. The corresponding bridge processes have tractable dynamics. This can be seen directly from the h-transform, since the gradient of the log-density is a linear function of the state, hence from (4) we have that the bridge process is also a linear SDE. This can be proved also from first principles working with the finite-dimensional distributions of the conditioned process, see for example Theorem 2 of Delyon and Hu (2006). In the simplest setup where $b = 0$ and σ is the identity matrix, the linear SDE is the Brownian motion and the bridge process conditioned upon $V_T = v$ is known as *Brownian bridge*, which solves the time-inhomogeneous SDE

$$\mathrm{d}V_s = \frac{v - V_s}{T - s}\mathrm{d}s + \mathrm{d}B_s\,, \tag{12}$$

and has macroscopic dynamics specified, for $0 < t_1 < t_2 < T$, as

$$V_{t_2} \mid V_{t_1} \sim N\left(V_{t_1} + \frac{t_2 - t_1}{T - t_1}(v - V_{t_1}),\, \frac{(t_2 - t_1)(T - t_2)}{T - t_1}\right)\,. \tag{13}$$

2.2 Importance sampling and identities

Importance sampling (IS) is a classic Monte Carlo technique for obtaining samples from a probability measure $\mathbb{P}$ using samples from another probability measure $\mathbb{Q}$, see for example Chapter 2.5 of Liu (2008) for an introduction. Mathematically it is based on the concept of *change of measure*. Suppose that $\mathbb{P}$ is *absolutely continuous* with respect to $\mathbb{Q}$ with Radon-Nikodym density $f(x) = \mathbb{P}(\mathrm{d}x)/\mathbb{Q}(\mathrm{d}x)$. Then, in its simplest form IS consists of constructing a set of *weighted particles* (x_i, w_i), $i = 1, \ldots, N$, where $x_i \sim \mathbb{Q}$, and $w_i = f(x_i)$. This set gives a Monte Carlo approximation of $\mathbb{P}$, in the sense that for suitably integrable functions g, we have that

$$\frac{\sum_{i=1}^{N} g(x_i) w_i}{N} . \tag{14}$$

is an unbiased and consistent estimator of

$$\mathrm{E}_{\mathbb{P}}[g] := \int g(x)\mathbb{P}(\mathrm{d}x) .$$

However, IS can be cast in much more general terms, an extension particularly attractive in the context of stochastic processes. First, note that in most applications f is known only up to a normalising constant, $f(x) = cf_u(x)$, where only f_u can be evaluated and

$$c = \mathrm{E}_{\mathbb{Q}}[f_u] . \tag{15}$$

The notion of a *properly weighted sample* (see for example Section 2.5.4 of Liu, 2008) refers to a set of weighted particles (x_i, w_i), where $x_i \sim \mathbb{Q}$ and w_i is an *unbiased estimator* of $f_u(x_i)$, that is

$$\mathrm{E}_{\mathbb{Q}}[w_i \mid x_i] = f_u(x_i) .$$

In this setup we have the fundamental equality for any integrable g

$$\mathrm{E}_{\mathbb{Q}}[gw] = \mathrm{E}_{\mathbb{P}}[g]\, \mathrm{E}_{\mathbb{Q}}[w] . \tag{16}$$

Rearranging the expression we find that a *consistent* estimator of $\mathrm{E}_{\mathbb{P}}[g]$ is given by

$$\frac{\sum_{i=1}^{N} g(x_i) w_i}{\sum_{i=1}^{N} w_i} . \tag{17}$$

When w_i is an unbiased estimator of $f(x_i)$ we have the option of using (14), thus yielding an unbiased estimator. However, (17) is a feasible estimator when c is unknown.

Although the first moment of w (under $\mathbb{Q}$) exists by construction, the same is not true for its second moment. Hence it is a minimal requirement of a "good proposal distribution $\mathbb{Q}$ that $\mathrm{E}_{\mathbb{Q}}[w^2] < \infty$. In this case, and using the Delta method for ratio of averages it can be shown that (17) is often preferable to (14) in a mean square error sense because the denominator acts effectively as a control variable. Therefore it might be preferable even when c is known. The same analysis leads to an interesting approximation for the variance of (17). Assume for simplicity that $\mathrm{E}_{\mathbb{Q}}[w] = 1$. Then exploiting the fact that

$$\mathrm{E}_{\mathbb{P}}[w] = \mathrm{var}_{\mathbb{Q}}[w] + 1\,,$$

using a further Taylor expansion for we obtain the following approximation for the variance of (17):

$$\frac{1}{N}\mathrm{var}_{\mathbb{P}}[g](1 + \mathrm{var}_{\mathbb{Q}}[w])\,. \tag{18}$$

Although this might be a poor approximation when the residual terms are significant, the expression motivates the notion of the *effective sample size (ESS)*, $1/(\mathrm{var}_{\mathbb{Q}}[w] + 1)$. This corresponds to an approximation of the ratio of variances of a Monte Carlo estimator of $\mathrm{E}_{\mathbb{P}}[g]$ based on independent samples from $\mathbb{P}$, and the IS estimator (17). The most appealing feature of the above approximation is that it does not depend on the function g, hence ESS can be used as a rough indication of the effectiveness of the IS approximation of $\mathbb{P}$. $N\times$ESS can be interpreted as the equivalent number of independent samples from $\mathbb{P}$. For more details see Section 2.5.3 of Liu (2008) and references therein.

The general framework where w is an unbiased estimator of the Radon-Nikodym derivative between $\mathbb{P}$ and $\mathbb{Q}$, opens various possibilities: constructing new Monte Carlo schemes (e.g Partial Rejection Control, see Section 2.6 of Liu, 2008), devising schemes whose computational complexity scales well with the number of particles N (e.g. the auxiliary particle filter of Pitt and Shephard, 1999), or applying IS in cases where even the computation of f_u is infeasible (e.g. the random weight IS for diffusions of Beskos et al., 2006; Fearnhead et al., 2008, which is also covered in detail in this article).

IS includes exact simulation as a special case when $\mathbb{Q} = \mathbb{P}$. Another special case is *rejection sampling* (RS), which assumes further that $f_u(x)$ is bounded in x by some calculable $K < \infty$. Then, if we accept each draw x_i with probability $f_u(x_i)/K$, the resulting sample (of random size) consists of independent draws from $\mathbb{P}$. This is a special case of the generalised IS where w_i is a binary 0-1 random variable taking the value 1 with probability $f_u(x_i)/K$.

The IS output can be used in estimating various normalising constants and density values involved in the costruction. It follows directly from the previous exposition that $c = \mathrm{E}_{\mathbb{Q}}[w]$. Moreover, note that

$$f(x) = \mathrm{E}_{\mathbb{Q}}[w \mid x]/\mathrm{E}_{\mathbb{Q}}[w]\,. \tag{19}$$

3 IS estimators based on bridge processes

Let V be a multivariate diffusion (1) observed at two consecutive time points $V_0 = u$, $V_T = v$, and consider the following two problems: a) (*imputation*) the design of efficient IS scheme for the corresponding diffusion bridge, and b) (*estimation*) the MC estimation of the corresponding transition density (2). In this section we consider these problems for a specific class of diffusions: those for which the diffusion coefficient is indepedent of V. In this context the methodology is much simpler and important developments have been made since the mid-80s. Furthermore, under additional structure exact simulation of diffusion bridges is feasible (see Section 5).

Before detailing the approach let us briefly discuss the restriction imposed by the assumption that the diffusion coefficient is independent of V. For scalar diffusions when $\sigma(\cdot,\cdot)$ is appropriately differentiable, by Itô's rule the transformation $V_s \to \eta(s, V_s) =: X_s$, where

$$\eta(s,u) = \int^u \frac{1}{\sigma(s,z)}\,\mathrm{d}z, \tag{20}$$

is any anti-derivative of $\sigma^{-1}(s,\cdot)$, yields a diffusion X with diffusion coefficient 1. A particle approximation of the diffusion bridge of X directly implies one for V and the transition densities of the two processes are linked by a change of variables formula. Therefore, the methodology of this section effectively covers all scalar diffusions, as well as a wide variety of multivariate processes used in applications. At the end of this section we discuss the limitations of the methodology based on bridge processes.

Let $\mathbb{P}_b$ be the law of the diffusion V on $[0,T]$ with $V_0 = u$ (abusing slighlty the notation set up in Section 2.1). Similarly, let $\mathbb{P}_0$ denote the law of the driftless process $\mathrm{d}V_s = \sigma(s)\mathrm{d}B_s$. Crucially, in this setting the driftless process is a linear SDE and $\mathbb{P}_0$ is a Gaussian measure. Additionally, let $p_{0,T}(u,v)$ and $\mathcal{G}_{0,T}(u,v)$ denote the transition densities of the two processes. Let $\mathbb{P}_b^*$ and $\mathbb{P}_0^*$ denote the laws of the corresponding diffusion bridges conditioned on $V_T = v$. As we discussed in Section 2.1, the conditioned driftless process is also a linear SDE.

We present a heuristic argument for deriving the density $\mathrm{d}\mathbb{P}_b^*/\mathrm{d}\mathbb{P}_0^*$. Consider the decomposition of the laws $\mathbb{P}_b$ and $\mathbb{P}_0$ into the marginal distributions at time T and the diffusion bridge laws conditioned on V_T. Then by a marginal-conditional decomposition we have that for a path V with $V_0 = u$,

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}(V)\ 1[V_T = v] = \frac{p_{0,T}(u,v)}{\mathcal{G}_{0,T}(u,v)}\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{P}_0^*}(V)\,. \tag{21}$$

The term on the left-hand side is given by the Cameron-Martin-Girsanov theorem (see Section 2.1). Hence, by a rearrangement we get the density between the diffusion bridge laws:

$$\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{P}_0^*}(V) = \frac{\mathcal{G}_{0,T}(u,v)}{p_{0,T}(u,v)} \exp\left\{\int_0^T h(s,V_s)^*\mathrm{d}B_s - \frac{1}{2}\int_0^T [h^*h](s,V_s)\mathrm{d}s\right\}\,, \tag{22}$$

where h solves $\sigma h = b$ (see Section 2.1), and B is Brownian motion.

Additional structure on b and σ can lead to further simplifications of (22). We emphasize the setting where σ is the identity matrix, the diffusion is time-homogenous and of *gradient-type*, i.e. there exists a field H such that $b(v) = \nabla_v H(v)$. When the function $\rho(v) \propto \exp\{H(v)/2\}$ is integrable, the diffusion is a reversible Markov process with ρ as the invariant density. In this setting, we can use Itô's rule to perform integration by parts in the exponent of (22) to eliminate the stochastic integral, and obtain

$$\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{P}_0^*}(V) = \frac{\mathcal{G}_{0,T}(u,v)}{p_{0,T}(u,v)} \exp\left\{H(v) - H(u) - \frac{1}{2}\int_0^T \left(||b(V_s)||^2 + \nabla^2 H(V_s)\right) \mathrm{d}s\right\} . \tag{23}$$

(22) forms the basis for a particle approximation of the law of $\mathbb{P}_b^*$ using proposals from $\mathbb{P}_0^*$. An idealized algorithm proceeds by first generating a linear SDE according to $\mathbb{P}_0^*$, and subsequently, by assigning weight according to (22). Note in particular that B in (22) is the Brownian motion driving the proposed linear bridge. The weights are known only up to a normalizing constant due to the presence of $p_{0,T}(u,v)$. However, as we saw in Section 2.2 this poses no serious complication in the application of IS. Note that $\mathcal{G}_{0,T}(u,v)$ is a Gaussian density which can be computed and be included explicitly in the weights, although this is not necessary for the IS.

Practically, we will have to simulate the proposed bridge at a finite collection of M times in $[0,T]$ and approximate the integrals in the weights by sums. This is an instance of the simulation-projection strategy outlined in Section 1. It introduces a bias in the MC approximations which is eliminated as $M \to \infty$. It is a subtle and largely unresolved issue how to distribute a fixed computational effort between M and N in order to minimize the MC variance of estimates of expectations of a class of test functions. However, a qualitative and asymptotic result is given in Stramer and Yan (2007) according to which one should choose $N = \mathcal{O}(M^2)$. In Section 5 we will see that in the more specific case of (23) the approximations can be avoided altogether and construct a properly weighted sample using unbiased estimators of the weights.

It follows directly from the general development of Section 2.2 that the diffusion transition density can be consistently estimated using a particle approximation of $\mathbb{P}_b^*$. From (15) it follows the key identity

$$p_{0,T}(u,v) = \mathcal{G}_{0,T}(u,v)\mathrm{E}_{\mathbb{P}_0^*}\left[\exp\left\{\int_0^T h(s,V_s)^*\mathrm{d}B_s - \frac{1}{2}\int_0^T [h^*h](s,V_s)\mathrm{d}s\right\}\right] \tag{24}$$

In the case where b is of gradient form we can correspondingly write

$$p_{0,T}(u,v) = \mathcal{G}_{0,T}(u,v)\exp\{H(v) - H(u)\}\mathrm{E}_{\mathbb{P}_0^*}\left[\exp\left\{-\int_0^T \phi(V_s)\mathrm{d}s\right\}\right] . \tag{25}$$

where $\phi(z) = (||b(z)||^2 + \nabla^2 H(z))/2$. Hence, the transition density is estimated by the average of the IS weights. It is at this stage where the explicit computation of the Gaussian density in the denominator of (22) becomes indispensable: if it were unknown we could only estimate the ratio of the two transition densities, but not $p_{0,T}(u,v)$.

Historical development

The expressions (22) and (24) have been derived several times in the literature with different motives. Remarkably, there is almost no cross-referencing among the papers which have derived the expressions. To our best knowledge, the expressions appear for the first time for scalar diffusions in the proof of Theorem 1 of Rogers (1985). The context of the Theorem is to establish smoothness of the transition density. Again for scalar diffusions the expressions appear in the proofs of Lemma 1 of Dacunha-Castelle and Florens-Zmirou (1986). The context of that paper is a quantification of the error in parameter estimates obtained using approximations of the transition density. Since both papers deal with scalar diffusions, they apply the integration by parts to get the simplified expression (23). More recently, Durham and Gallant (2002) working in a projection-simulation paradigm, derive effectively an IS for $\mathbb{P}_b^*$ and an estimator of $p_{0,T}(u,v)$, which in the case of constant diffusion coefficient are discretizations of (22) and (24) (see also Section 4 below). The context here is MC estimation of diffusion models. Since the authors work in a time-discretized framework from the beginning, the possibility to perform integration by parts when possible, is not at all considered. Nicolau (2002) uses the Dacunha-Castelle and Florens-Zmirou (1986) expression for the transition density as a basis for MC estimation using approximation of the weights based on M intermediate points. Beskos et al. (2006) used (23) as a starting point for the exact simulation of diffusions and (24) as a basis for unbiased estimation of the transition density (see also Section 5). Finally, Delyon and Hu (2006) state (22) as Theorem 2 and prove it for mutivariate processes.

Limitations of the methodology

The outline of the methodology we have described in this section for IS approximation of $\mathbb{P}_b^*$ is to find a probability measure $\mathbb{P}_0$ which is absolutely continuous with respect to the unconditional measure $\mathbb{P}_b$, and *probabilistically condition* the former on the same event that the latter is conditioned upon. Hence, the proposed random variables are indeed *bridge processes*. The same development can be carried out even when σ depends on the state variable. In the more general setup $\mathbb{P}_0$ is the law of $\mathrm{d}V_s = \sigma(s, V_s)\mathrm{d}B_s$, which is now a non-linear diffusion. Hence the corresponding bridge process will be typically intractable (see Section 2.1) and the IS practically infeasible. Therefore, the methodology of this section applies only to diffusions which can be transformed to have state-independent diffusion coefficient.

For multivariate diffusions with V-dependent volatility the generalized version of transformation (20) involves the solution of an appropriate vector differential equation which is often intractable or insolvable, see for example Aït-Sahalia (2008). A very popular class of models which have state-dependent volatility are stochastic volatility models employed in financial econometrics. The methodology of space transformation can be generalised to include time-change methodology to allow models like stochastic volatility models to be addressed, (see Kalogeropoulos et al., 2009).

Summarising, constructing valid and easy to simulate proposals by *conditioning* is difficult when we deal with multivariate processes with state-dependent diffusion coefficient. Instead, the next section considers a different possibility where the proposals are generated by a process which is explicitly constructed to hit the desired end-point at time T. We call such processes *guided*.

4 IS estimators based on guided processes

In this section we consider the same two problems described in the beginning of Section 3 but for processes with state-dependent diffusion coefficient. We consider IS approximation of the law of the target diffusion bridge, $\mathbb{P}_b^*$, using appropriate diffusion processes as proposals. The design of such proposals is guided by the SDE of the target bridge, given in (4), and the Cameron-Martin-Girsanov theorem. Hence, the diffusion coefficient of any valid proposal has to be precisely $\sigma(s,v)$, and the drift has to be such that it forces the process to hit the value v at time T almost surely. The following processes are natural candidates under these considerations:

$$\begin{aligned} \text{[G1]} \quad \mathrm{d}V_s &= -\frac{V_s - v}{T - s}\,\mathrm{d}s + \sigma(s, V_s)\,\mathrm{d}B_s && (26) \\ \text{[G2]} \quad \mathrm{d}V_s &= -\frac{V_s - v}{T - s}\,\mathrm{d}s + b(s, V_s)\,\mathrm{d}s + \sigma(s, V_s)\mathrm{d}B_s && (27) \end{aligned}$$

Note that the drift of [G1] ("G" stands for "Guided") is precisely the one of the Brownian bridge (12); the one of [G2] mimics the structure of the drift of the target bridge process (4) but substitutes the intractable term in the drift by the Brownian bridge drift. Let $\mathbb{Q}_{G1}$ and $\mathbb{Q}_{G2}$ denote the laws of [G1] and [G2] correspondingly. We will use [G] to refer to a generic guided process.

The mathematical argument which yields the IS is formally presented in Section 4 of Delyon and Hu (2006). The construction requires for tractability that σ is invertible so that we can work with (7) and introduce explicitly V (instead of the driving B) in the weights. To simplify the exposition, we present the argument when $d = 1$; the formulae extend naturally to the multidimensional case, under the same assumptions.

We define for any z and $s \leq T$,

$$A(s, z) = (\sigma(s, z))^{-2}\,.$$

Up to any time $t < T$, $\mathbb{Q}_{G2}|_t$ (resp. $\mathbb{Q}_{G1}|_t$) is absolutely continuous with respect to $\mathbb{P}_b^*|_t$, and we can apply the Cameron-Martin-Girsanov theorem (7). The resulting likelihood ratio, although of expected value 1, it converges almost surely to 0 as $t \to T$. To identify the leading term which drives the weights to 0 (denoted ψ_t below) we apply an integration by parts to the stochastic integral in the exponent. Let us define the following functionals

$$\begin{aligned}
\psi_t &= \exp\left\{-\frac{1}{2(T-t)}(V_t-v)^2 A(t,V_t)\right\} \\
C_t &= \frac{1}{(T-t)^{1/2}} \\
\log(\phi_t^{G1}) &= -\int_0^t \frac{(V_s-v)}{T-s} A(s,V_s) b(s,V_s) \mathrm{d}s - \int_0^t \frac{(V_s-v)^2}{2(T-s)} \diamond \mathrm{d}A(s,V_s) \\
\log(\phi_t^{G2}) &= \int_0^t b(s,V_s) A(s,V_s) \mathrm{d}V_s - \frac{1}{2}\int_0^t b(s,V_s)^2 A(s,V_s) \mathrm{d}s - \int_0^t \frac{(V_s-v)^2}{2(T-s)} \diamond \mathrm{d}A(s,V_s)
\end{aligned}$$

where the $\diamond$-stochastic integral is understood as the limit of approximating sums where the integrand is evaluated at the right-hand time-points of each sub-interval (as opposed to the left-hand in the definition of the Itô stochastic integral). Then, we have that

$$\begin{aligned}
\left.\frac{\mathrm{d}\mathbb{Q}_{G2}}{\mathrm{d}\mathbb{P}_b}\right|_t &= \sqrt{T}\exp\left\{\frac{(u-v)^2 A(0,u)}{2T}\right\} C_t \psi_t / \phi_t^{G2} \\
\left.\frac{\mathrm{d}\mathbb{Q}_{G1}}{\mathrm{d}\mathbb{P}_b}\right|_t &= \sqrt{T}\exp\left\{\frac{(u-v)^2 A(0,u)}{2T}\right\} C_t \psi_t / \phi_t^{G1} .
\end{aligned}$$

Therefore, for any measurable (with respect to the filtration of V up to t) non-negative function f_t, we have that

$$\begin{aligned}
\mathrm{E}_{\mathbb{P}_b}[f_t(V)\psi_t] &= C_t^{-1} T^{-1/2} \exp\left\{-\frac{(u-v)^2 A(0,u)}{2T}\right\} \mathrm{E}_{\mathbb{Q}_{G2}}[f_t(V)\phi_t^{G2}] \\
&= C_t^{-1} T^{-1/2} \exp\left\{-\frac{(u-v)^2 A(0,u)}{2T}\right\} \mathrm{E}_{\mathbb{Q}_{G1}}[f_t(V)\phi_t^{G1}] \\
\mathrm{E}_{\mathbb{P}_b}[\psi_t] &= C_t^{-1} T^{-1/2} \exp\left\{-\frac{(u-v)^2 A(0,u)}{2T}\right\} \mathrm{E}_{\mathbb{Q}_{G2}}[\phi_t^{G2}] \\
&= C_t^{-1} T^{-1/2} \exp\left\{-\frac{(u-v)^2 A(0,u)}{2T}\right\} \mathrm{E}_{\mathbb{Q}_{G1}}[\phi_t^{G1}] ,
\end{aligned}$$

where the expressions for $\mathrm{E}_{\mathbb{P}_b}[\psi_t]$ is obtained from the first expression with $f_t = 1$.

Hence, we derive the key equality, that for any positive measurable f_t,

$$\frac{\mathrm{E}_{\mathbb{P}_b}[f_t(V)\psi_t]}{\mathrm{E}_{\mathbb{P}_b}[\psi_t]} = \frac{\mathrm{E}_{\mathbb{Q}_{G2}}[f_t(V)\phi_t^{G2}]}{\mathrm{E}_{\mathbb{Q}_{G2}}[\phi_t^{G2}]} = \frac{\mathrm{E}_{\mathbb{Q}_{G1}}[f_t(V)\phi_t^{G1}]}{\mathrm{E}_{\mathbb{Q}_{G1}}[\phi_t^{G1}]},.$$

The final part of the argument consists of taking the limit $t \to T$ on each part of the previous equality (this requires a careful non-trivial technical argument, see proof of Theorem 5 and related Lemmas of Delyon and Hu (2006)). The limit on the left hand side converges to the regular conditional expectation $\mathrm{E}_{\mathbb{P}_b^*}[f_T(V)] = \mathrm{E}_{\mathbb{P}_b}[f_T(V) \mid V_T = v]$; intuitively this can be verified by the form of ψ_t given above. The other two terms converge to $\mathrm{E}_{\mathbb{Q}_{G2}}[f_T(V)\phi_T^{G2}]/\mathrm{E}_{\mathbb{Q}_{G2}}[\phi_T^{G2}]$ and $\mathrm{E}_{\mathbb{Q}_{G2}}[f_T(V)\phi_T^{G1}]/\mathrm{E}_{\mathbb{Q}_{G1}}[\phi_T^{G1}]$, respectively. Therefore, we have that

$$\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{Q}_{G2}}(V) = \frac{\phi_T^{G2}}{\mathrm{E}_{\mathbb{Q}_{G2}}[\phi_T^{G2}]} \tag{28}$$

$$\frac{\mathrm{d}\mathbb{P}_b^*}{\mathrm{d}\mathbb{Q}_{G1}}(V) = \frac{\phi_T^{G1}}{\mathrm{E}_{\mathbb{Q}_{G1}}[\phi_T^{G1}]} \tag{29}$$

where the denominators on the right-hand side in each expression are normalising constants. These two expressions are all is needed for the IS approximation of the diffusion bridge. Practically, as in Section 3, we will have to simulate the proposed bridge at a finite collection of M times in $[0, T]$ and approximate the integrals in the weights by sums, which introduces a bias in the MC approximations which is eliminated as $M \to \infty$.

We now address the problem of deriving a transition density identity, as we did in (24). To our best knowledge, this is the first time that such an expression appears in the literature. Note, however, that our argument is informal and certain technical conditions (outside the scope of this article) will have to be imposed for a formal derivation. Working with guided processes, this derivation is much less immediate than in Section 3.

Since ψ_t is a function of V_t only, we have that

$$\begin{aligned}
\mathrm{E}_{\mathbb{P}_b}[C_t\psi_t] &= \int \frac{1}{\sqrt{T-t}} \exp\left\{-\frac{1}{2(T-t)}(w-v)^2 A(t,w)\right\} p_{0,t}(u,w)\mathrm{d}w \\
&= \int \exp\left\{-\frac{1}{2}z^2 A(t, z\sqrt{T-t}+v)\right\} p_{0,t}(u, z\sqrt{T-t}+v)\mathrm{d}z \\
&\to_{t\to T} \quad p_{0,T}(u,v) \int \exp\left\{-\frac{1}{2}z^2 A(T,v)\right\} \mathrm{d}z = \frac{\sqrt{2\pi}\, p_{0,T}(u,v)}{\sqrt{A(T,v)}}
\end{aligned}$$

where taking the limit we have used dominated convergence (which clearly requires certain assumptions). We can use this epxression, together with the identities which link $\mathrm{E}_{\mathbb{P}_b}[C_t\psi_t]$ with $\mathrm{E}_{\mathbb{Q}_{Gi}}[\phi_t^{Gi}]$, $i = 1, 2$ given above, and the fact that the latter converge as $t \to T$ to $\mathrm{E}_{\mathbb{Q}_{Gi}}[\phi_T^{Gi}]$, $i = 1, 2$ (which is shown in Delyon and Hu (2006)), to establish the fundamental identity

$$p_{0,T}(u,v) = \sqrt{\frac{A(T,v)}{2\pi T}} \exp\left\{-\frac{(u-v)^2 A(0,u)}{2T}\right\} \mathrm{E}_{\mathbb{Q}_{Gi}}[\phi_T^{Gi}]\,, \; i = 1, 2, . \tag{30}$$

Therefore, given the IS output the transition density can be estimated.

Connections to the literature and to Section 3

The first major contribution to IS in this context was made in the seminal article of Durham and Gallant (2002) in the context of estimating the transition density of non-linear diffusions for statistical inference. They took, however, a projection-simulation approach, they first discretized the unobserved paths and then considered discrete-time processes as proposals. They suggest two different discrete-time processes as proposals, the so-called "Brownian bridge" proposal and the "modified Brownian bridge". They both are inspired by (and intend to be a type of discretization of) (26). Indeed, their "Brownian bridge" proposal is precisely a first-order Euler approximation of (26). The Euler approximation is not very appealing here since it is unable to capture the inhomogeneity in the variance of the transition distribution of (26). To see this more clearly, consider the simplified case where $\sigma = 1$ and contrast (12) with (13) when $t_2 - t_1$ is small. The Euler approximation suggests constant variance for the time-increments of the process, which is a very poor approximation when $t \approx T$. To mitigate against this lack of heteroscedasticity, Durham and Gallant (2002) use Bayes' theorem together with heuristic approximations to find a better aproximation to the transition density of (26). The process with the new dynamics is termed "modified Brownian bridge", and it corresponds to the exact solution of the Brownian bridge SDE when $\sigma = 1$. Generally, due to various approximations at various stages the connection between IS for paths and estimation of the transition density is not particularly clear in their paper.

It is important to observe that the samplers and identities in this section become precisely those of Section 3 when σ is constant. An interesting deviation, is the use of (27) when σ is constant, which does not correspond to the setup of Section 3, and has the effect of making non-linear the proposal process (hence exact skeletons cannot be simulated in this case) but removes the stochastic integral from the weights (which typically has as a variance reduction effect).

As a final remark note that when possible it is advisable to transform the diffusion to have unit volatility. Apart from facilating the methodology Section 3 the tansformation has been empirically shown to be a good variance reduction technique even for schemes which do not require it (see for example the discussion in Durham and Gallant, 2002).

5 Unbiased Monte Carlo for diffusions

In many cases (including most all one-dimensional diffusions with sufficiently smooth coefficients) the need to use a fine discretisation for the diffusion sample path (and the associated approximation error) can be completely removed. This section will very briefly describe some of the basic ideas behind this approach, though for detailed account the reader is referred to Beskos et al. (2006); Fearnhead et al. (2008). The methodology here is closely related to allied exact simulation algorithms for diffusions as described in Beskos and Roberts (2005); Beskos et al. (2004, 2005b).

For simplicity we shall focus on the problem of estimating (unbiasedly) the diffusion transition density. The use of this approach in Monte Carlo maximum likelihood and related likelihood inference methodology is described in detail in Beskos et al. (2006). We shall assume that the diffusion can be reduced to unit diffusion coefficient, and that the drift b can be expressed in gradient form $b = \nabla H$. We can therefore express the transition density according to 25.

Here we describe the so-called *generalised Poisson estimator* for estimating $p_{0,T}(u,v)$. A simple Taylor expansion of 25 gives

$$p_{0,T}(u,v) = \mathcal{G}_{0,T}(u,v)\exp\{H(v) - H(u)\}\mathrm{E}_{\mathbb{P}_0^*}\left[\sum_{\kappa=0}^{\infty} \frac{\left(\int_0^T \phi(V_s)\right)^{\kappa}}{\kappa!}\right]. \tag{31}$$

A simple observation we can make is that for an arbitrary function $g(\cdot)$, $\mathrm{E}(\int_0^T g(X_s)ds)$ is readily estimated unbiasedly by $Tg(_U)$ where $U \sim U(0,T)$. This idea is easily generalised to consider $\mathrm{E}((\int_0^T g(X_s)ds)^{\kappa})$ for arbitrary positive integer κ. In this case the unbiased estimator is just $T^{\kappa}\prod_{i=1}^{\kappa} g(X_{U_i})$ where $\{U_i\}$ denote an independent collection of $U(0,T)$ variables.

Therefore, letting $\{q_i\}$ denote positive probabilities for all non-negative integers i, and unbiased estimator for $p_{0,T}(u,v)$ is given for arbitrary constant c by

$$\hat{p}_{0,T}(u,v) = \mathcal{G}_{0,T}(u,v)\exp\{H(v) - H(u) - cT\}T^I\prod_{i=1}^{I}(c - \phi(V_{U_i}))q_I^{-1} \tag{32}$$

where $I \sim q$. The choice of the importance proposal q is critical in determining the efficiency of the estimator $\hat{p}_{0,T}(u,v)$, and this is discussed in more detail in Beskos et al. (2006); Fearnhead et al. (2008).

For the purposes of parameter estimation, it is critical to be able to obtain density estimates simultaneously for a collection of plausible parameter values. This is one important advantage of the form of the estimator in 32 since the estimator can be applied simultaneously to provide unbiased estimators of densities at a continuum of parameter values in such a way that the estimated likelihood surface is itself continuous (and of course unbiased for each parameter choice). As well as being of practical use, these properties are indispensible for proving consistency of Monte Carlo MLEs for large sample sizes (Beskos et al., 2005a)

Acknowledgements

The first author would like to acknowledge financial support from the Spanish Goverment via grant MTM2008-06660 and the “Ramon y Cajal” fellowship. We would like also to think Jaya Bishwal and Giorgos Sermaidis for helpful suggestions.

Appendix 1: typical problems of the projection-simulation paradigm in MC for diffusions

In this article we have advocated a simulation-projection paradigm, that is designing Monte Carlo methods on the path space which are then, if necessary, discretized for practical implementation. Apart from the transparency of the resulting methods, the motivation for adopting this paradigm is also due to typical problems faced by the projection-simulation alternative. In this Appendix we mention two typical problematic cases. The first concerns the estimation of the transition density by Monte Carlo, and the the simultaneous parameters estimation and imputation of unobserved paths using the Gibbs sampler. A common characteristic in both is that decrease in approximation bias comes with an increase in Monte Carlo variance. For the sake of presentation we only consider scalar homogeneous diffusions.

The problem of estimating the transition density by Monte Carlo and use the approximation for likelihood inference for unknown parameters was first considered by Pedersen (1995). Using the Chapman-Kolmogorov equation and Euler approximation he obtained

$$\begin{aligned} p_{0,T}(u,v) &= \mathrm{E}_{\mathbb{P}_b}\left[p_{t,T}(V_t,v)\right] \\ &\approx \mathrm{E}_{\mathbb{P}_b}\left[C_t\psi_t\frac{1}{\sqrt{2\pi}}A(t,V_t)^{1/2}\exp\left\{-\frac{1}{2}A(t,V_t)(b(t,V_t)^2(T-t)+2(v-V_t)b(t,V_t))\right\}\right] \end{aligned} \tag{33}$$

with the definitions as in Section 4. This suggest an IS approximation where we generate (unconditionally) paths up to time $t < T$ and associate weights to each path given by

$$\psi_t A(t,V_t)^{1/2}\exp\left\{-\frac{1}{2}A(t,V_t)(b(t,V_t)^2(T-t)+2(v-V_t)b(t,V_t))\right\}.$$

Due to the Euler approximation on $[t,T]$ the weights have a bias which is eliminated as $t \to T$. On the other hand, the leading term in the weights for $t \approx T$ is ψ_t, thus the variance of the weights tends to infinity as $t \to T$. (There is of course additional potential bias in simulating V_t using a discretization method, this however can be eliminated with increasing Monte Carlo effort without inflating the variance of the weights). The approach we expose in Sections 3 and 4 is designed to overcome this problem.

The problem of Bayesian inference for unknown parameters in the drift and the volatility of the SDE and the simultaneous imputation of unobserved paths for discretely observed diffusions was originally considered by Elerian et al. (2001); Eraker (2001); Roberts and Stramer (2001) (and remains a topic of active research). The first two articles work in a projection-simulation framework, hence the unobserved path between each pair of observations (i.e each diffusion bridge) is approximated by a skeleton of, M say, points. The joint distribution of the augmented dataset can be approximated using for example the Euler scheme (which gets increasingly accurate as M increases). This is effectively equivalent to using a Riemmann approximation to the continuous-time likelihood (7). Therefore, we deal with a missing data problem where, given additional data (the imputed values in-between the observations) the likelihood is available, although here the missing data (for each pair of obsevrations) are in principle infinite-dimensional and are approximated by an M-dimensional vector. Hence the computations are subject to a model-approximation bias which is eliminated in the limit $M \to \infty$. The *Gibbs sampler* is a popular computational tool for parameter estimation and simultaneous imputation of missing data in such a context. It consists of iterative simulation of missing data given the observed data and current values of prameters, and the simulation of the parameters according to their posterior distribution conditionally on the augmented dataset.

There are two main challenges in designing a Gibbs sampler for discretely observed diffusions: how to efficiently simulate the M intermediate points given the endpoints for each pair of observations, and how to reduce the dependence between the missing data and the parameters. As far as the first problem is concerned, note that it is directly related to the diffusion bridge simulation, and it is best understood thinking of the simulation in the infinite-dimensional space. For diffusions which can be transformed to have unit volatility (as in Section 3) Roberts and Stramer (2001) describe a Markov chain Monte Carlo (MCMC) scheme which uses global moves on the path space, an approach very closely related to that described in Section 3. More recently, global moves MCMC using the processes discussed in Section 4 as proposals has been considered by Golightly and Wilkinson (2008); Chib et al. (2004). For local moves MCMC designed on the path space see for example Beskos et al. (2008).

However, it is the second challenge we wish to emphasize in this section, i.e. the dependence between the imputed data and the parameters. Strong posterior dependence between missing data and parameters is known to be the principal reason for slow convergence of the Gibbs sampler and results in high variance of the estimates based on its output (see for example Papaspiliopoulos et al., 2007). The dependence between imputed data and parameters in this application can only be understood by considering a Gibbs sampler on the infinite-dimensional space, i.e the product space of parameters and diffusion bridges. This approach was adopted in Roberts and Stramer (2001) where it was noticed that due to the quadratic variation identity (5) there is complete dependence between the missing paths and any parameters involved in the volatility. Hence, an idealized algorithm ($M = \infty$) would be completely reducible, whereas in practical applications where M is finite we observe that decreasing the bias (increasing M) causes an increase of the mixing time of the algorithm and a corresponding increase in the variance of estimates based on its output. A solution to this problem is given in Roberts and Stramer (2001) by appropriate transformations in the path space which break down this dependence. It turns out that the strong dependence between parameters and unobserved processes is very common in many hierarchical models and a generic methodology for reducing it, which includes the one considered in Roberts and Stramer (2001), is known as *non-centred parametrisations*, see Papaspiliopoulos et al. (2003, 2007).

Appendix 2: Gaussian change of measure

The concept of change of measure is very central to the approaches we have treated in this article. The aim of this section is to give a simplified presentation of the change of measure between two Gaussian laws, and to the various ways this result might be put in use. It is easy to see the correspondence between the expressions we obtain here and those of Section 2.1, but the greatly simplified context of this section has the educational value of pointing out some of the main elements of the construction, which can be understood without knowledge of stochastic calculus.

Let $(\Omega, \mathcal{F})$ be a measure space with elements $\omega \in \Omega$, $B : \Omega \to R^m$ a random variable on that space, let σ be a $d \times m$ matrix, $\Gamma = \sigma\sigma^*$, a, b, be $d \times 1$ vectors, and define a random variable V via the equation

$$V(\omega) = b + \sigma B(\omega)\,.$$

Let $\mathbb{R}_b$ be the probability measure on $(\Omega, \mathcal{F})$ such that B is a standard Gaussian vector. Therefore, under this measure V is a Gaussian vector with mean b (hence the indexing of the measure by b). Assume now that we can find a $m \times 1$ vector h which solves the equation

$$\sigma h = (b - a)\,, \tag{34}$$

and define $\hat{B}(\omega) = B(\omega) + h$. Thus, we have the alternative representation

$$V(\omega) = a + \sigma \hat{B}(\omega)\,,$$

which follows directly from the definitions of V and h. Let $\mathbb{R}_a$ be the measure defined by its density with respect to $\mathbb{R}_b$,

$$\frac{\mathrm{d}\mathbb{R}_a}{\mathrm{d}\mathbb{R}_b}(\omega) = \exp\left\{-h^* B(\omega) - h^* h/2\right\}\,, \tag{35}$$

which is well-defined since the right-hand side has finite expectation with respect to $\mathbb{R}_b$. Notice that under this new measure, $\hat{B}$ is a standard Gaussian vector. To see this, notice that for any Borel set $A \subset R^m$,

$$\begin{aligned}
\mathbb{R}_a[\hat{B} \in A] &= \int_{\{\omega : \hat{B}(\omega) \in A\}} \exp\left\{-u^* B(\omega) - u^* u/2\right\} \mathrm{d}\mathbb{R}[\omega] \\
&= \int_{\{y : y+u \in A\}} \exp\left\{-u^* y - u^* u/2 - y^* y/2\right\} (2\pi)^{-m/2} \mathrm{d}y \\
&= \int_A e^{-v^* v/2} (2\pi)^{-m/2} \mathrm{d}v\,,
\end{aligned}$$

where the last equality follows from a change of variables.

Notice that directly from (35) we have

$$\frac{d\mathbb{R}_b}{d\mathbb{R}_a}(\omega) = \exp\{h^* B(\omega) + h^* h/2\} = \exp\{h^* \hat{B} - h^* h/2\}\,. \tag{36}$$

Let E_b and E_a denote expectations with respect to $\mathbb{R}_b$ and $\mathbb{R}_a$ respectively. Thus, for any measurable $\mathbb{R}_b$-integrable function f defined on R^d,

$$\mathrm{E}_b[f(V)] \quad = \quad \mathrm{E}_a\left[f(V)\exp\{h^* B + h^* h/2\}\right] = \mathrm{E}_a\left[f(V)\exp\{h^* \hat{B} - h^* h/2\}\right]\,.$$

Let X be another random variable, defined as $X(\omega) = a + \sigma B(\omega)$. Since under $\mathbb{R}_a$, the pair $(V, \hat{B})$ has the same law as the pair (X, B) under $\mathbb{R}_b$, we have that

$$\mathrm{E}_b[f(V)] = \mathrm{E}_b[f(X)\exp\{h^* B - h^* h/2\}]\,.$$

If further σ is invertible we get

$$\mathrm{E}_b[f(V)] = \mathrm{E}_b\left[f(X)\exp\left\{(b-a)^*\Gamma^{-1}X - \frac{1}{2}(b-a)^*\Gamma^{-1}(b+a)\right\}\right]\,. \tag{37}$$

Let $\mathbb{P}_b$ and $\mathbb{P}_a$ be the law of V implied by $\mathbb{R}_b$ and $\mathbb{R}_a$ respectively. Then, assuming that σ is invertible and taking $\alpha = 0$, we can obtain from the previous expression the likelihood ratio between the hypotheses that V has mean b against that it has mean 0, but a Gaussian distribution with covariance Γ in both cases. Therefore, we get the likelihood function for estimating b on the basis of observed data V, while treating Γ as known:

$$L(b) = \frac{d\mathbb{P}_b}{d\mathbb{P}_0}(V) = \exp\left\{b^*\Gamma^{-1}V - \frac{1}{2}b^*\Gamma^{-1}b\right\}\,. \tag{38}$$

It is interesting to consider the cases where (34) has many or no solutions. We will do so by looking at two characteristic examples. We first consider the case where (34) has multiple solutions and take $d = 1,\ m = 2,\ \sigma = (1, 1)$, in which case (34) has infinite solutions. Notice that in this case there are more sources of randomness than observed variables. To simplify matters (and without loss of generality) we take $a = 0$. Then, for any $\phi \in R$, $u = (\phi, b - \phi)^*$ solves (34), and the measure $\mathbb{R}_0^\phi$ defined by (35), makes $\hat{B}$ a standard Gaussian vector. Then, writing $B = (B_1, B_2)$, the importance weights in (37) become

$$\exp\{(b-\phi)B_1 + \phi B_2 - \phi^2 - b^2/2 + \phi b\}\,. \tag{39}$$

Direct calculation verifies that the change of measure in (37) holds for any ϕ; it is instructive to do directly the calculations using a change of variables and check that the right hand side of (37) does not depend on ϕ. Additionally, using the moment generating function of the Gaussian distribution, one can verify that the expected value of the importance weights (39) under $\mathbb{R}_0^\phi$ is 1. However, the second moment of the importance weights is $\exp\{2(\phi-b/2)^2+b^2/2\}$, which is minimized for $\phi = b/2$. Additionally, we can re-express (39) in terms of V as

$$\exp\{(b-\phi)V + (2\phi - b)\hat{B}_1 - \phi^2 - b^2/2 + \phi b\}\,.$$

In a statistical application only V will be observed whereas $\hat{B}_1$ will be unobserved, therefore we cannot use the expression directly to estimate b. Notice that for $\phi = b/2$ the $\hat{B}_1$ terms cancels out from the density.

We now consider the case where (34) has no solution. An example of that is produced under the setting $d = 2,\ m = 1,\ \sigma = (0,1)^*$. Writing $V = (V_1, V_2)^*$ and $a = (a_1, a_2)$, $b = (b_1, b_2)^*$, notice the example implies that $V_1 = b_1$. Therefore, it is expected that $\mathbb{R}_b$ will be mutually singular with any measure which implies that $V_1 = a_1$, if $b_1 \neq a_1$. However, notice that (34) can be solved by $u = b_2 - a_2$ provided that $a_1 = b_1$. Then, (35)-(37) hold. Moreover, defining $\mathbb{P}_{(b_1,b_2)}$ and $\mathbb{P}_{(b_1,0)}$ analogously as before, and noticing that $B = V_2 - b_2$, we have the following likelihood ratio which can be used for the estimation of b_2:

$$L(b_2) = \frac{\mathrm{d}\mathbb{P}_{(b_1,b_2)}}{\mathrm{d}\mathbb{P}_{(b_1,0)}}(V) = \exp\left\{b_2 V - \frac{1}{2}b_2^2\right\}\,.$$

References

Aït-Sahalia, Y. (2008) Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics*, **36**, 906–937.

Arnold, L. (1998) *Random dynamical systems.* Springer Monographs in Mathematics. Berlin: Springer-Verlag.

Bar-Shalom, Y., Kirubarajan, T. and Li, X.-R. (2002) *Estimation with Applications to Tracking and Navigation.* New York, NY, USA: John Wiley & Sons, Inc.

Bergstrom, A. R. (1990) *Continuous Time Econometric Modelling.* Oxford University Press.

Beskos, A., Papaspiliopoulos, O. and Roberts, G. O. (2004) Retrospective exact simulation of diffusion sample paths with applications. Submitted, available from http://www.maths.lancs.ac.uk/∼papaspil/research.html.

— (2005a) Monte carlo maximum likelihood estimation for discretely observed diffusion processes. Submitted.

— (2005b) A new factorisation of diffusion measure with view towards simulation. In progress.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and efficient likelihood–based inference for discretely observed diffusions (with Discussion). *J. Roy. Statist. Soc. Ser. B*, **68**, 333–82.

Beskos, A. and Roberts, G. O. (2005) Exact simulation of diffusions. *Ann. Appl. Probab.*, **15**. To appear.

Beskos, A., Roberts, G. O. and Stuart, A. M. (2008) Optimal scalings for local metropolis-hastings chains on non-product targets in high dimensions. *Annals of Applied Probability*, **to appear**.

Chan, K., Karolyi, A. G., Longstaff, F. A. and Sanders, A. B. (1992) An empirical comparison of alternative models of the short-term interest rate. *J. Finance*, **47**, 1209–1227.

Chib, S., Shephard, N. and Pitt, M. (2004) Likelihood based inference for diffusion driven models. Available from http://ideas.repec.org/p/sbs/wpsefe/2004fe17.html.

Cox, J. C., Ingersoll, Jr., J. E. and Ross, S. A. (1985) A theory of the term structure of interest rates. *Econometrica*, **53**, 385–407.

Dacunha-Castelle, D. and Florens-Zmirou, D. (1986) Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, **19**, 263–284.

Delyon, B. and Hu, Y. (2006) Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Process. Appl.*, **116**, 1660–1675.

Durham, G. B. and Gallant, A. R. (2002) Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *J. Bus. Econom. Statist.*, **20**, 297–338. With comments and a reply by the authors.

Elerian, O., Chib, S. and Shephard, N. (2001) Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, **69**, 959–993.

Eraker, B. (2001) MCMC analysis of diffusion models with application to finance. *J. Bus. Econom. Statist.*, **19**, 177–191.

Fearnhead, P., Papaspiliopoulos, O. and Roberts, G. O. (2008) Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 755–777.

Givon, D., Kupferman, R. and Stuart, A. (2004) Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, **17**, R55–R127.

Gobet, E. (2002) LAN property for ergodic diffusions with discrete observations. *Ann. Inst. H. Poincaré Probab. Statist.*, **38**, 711–737.

Golightly, A. and Wilkinson, D. J. (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics and Data Analysis*, **52**, 1674–1693.

Hurn, A.S. Jeisman, J. I. and Lindsay, K. (2007) Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics*, **5**, 390–455.

Kalogeropoulos, K., Roberts, G. O. and Dellaportas, P. (2009) Inference for stochastic volatility models using time change transformations. in revision - annals of statistics. *submitted*.

van Kampen, N. G. (1981) *Stochastic processes in physics and chemistry*. Amsterdam: North-Holland Publishing Co. Lecture Notes in Mathematics, 888.

Kessler, M. (1997) Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.*, **24**, 211–229.

Kimura, M. and Ohta, T. (1971) *Theoretical aspects of population genetics*. Princeton University Press.

Kloeden, P. and Platen, E. (1995) *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag.

Kutoyants, Y. A. (2004) *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. London: Springer-Verlag London Ltd.

Liu, J. S. (2008) *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. New York: Springer.

Maruyama, G. (1955) Continuous Markov processes and stochastic equations. *Rend. Circ. Mat. Palermo (2)*, **4**, 48–90.

McAdams, H. and Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA*, **94**, 814–819.

Merton, R. C. (1971) Optimum consumption and portfolio rules in a continuous-time model. *J. Econom. Theory*, **3**, 373–413.

Nicolau, J. (2002) A new technique for simulating the likelihood of stochastic differential equations. *Econom. J.*, **5**, 91–103.

Øksendal, B. K. (1998) *Stochastic Differential Equations: An Introduction With Applications*. Springer-Verlag.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2003) Non-centered parameterizations for hierarchical models and data augmentation. In *Bayesian statistics, 7 (Tenerife, 2002)*, 307–326. New York: Oxford Univ. Press. With a discussion by Alan E. Gelfand, Ole F. Christensen and Darren J. Wilkinson, and a reply by the authors.

— (2007) A general framework for the parametrization of hierarchical models. *Statist. Sci.*, **22**, 59–73.

Pedersen, A. R. (1995) Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, **1**, 257–279.

Pitt, M. K. and Shephard, N. (1999) Filtering via simulation: auxiliary particle filters. *J. Amer. Statist. Assoc.*, **94**, 590–599.

Prakasa Rao, B. L. S. (1999) *Statistical inference for diffusion type processes*, vol. 8 of *Kendall's Library of Statistics*. London: Edward Arnold.

Roberts, G. O. and Stramer, O. (2001) On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, **88**, 603–621.

Rogers, L. C. G. (1985) Smooth transition densities for one-dimensional diffusions. *Bull. London Math. Soc.*, **17**, 157–161.

Rogers, L. C. G. and Williams, D. (2000) *Diffusions, Markov processes, and martingales. Vol. 1.* Cambridge Mathematical Library. Cambridge: Cambridge University Press. Foundations, Reprint of the second (1994) edition.

Stramer, O. and Yan, J. (2007) Asymptotics of an efficient Monte Carlo estimation for the transition density of diffusion processes. *Methodol. Comput. Appl. Probab.*, **9**, 483–496.

Stuart, A. M. and Pokern, Y. (2009) Chapter for semstat.

Sundaresan, S. M. (2000) Continuous-time methods in finance: A review and an assessment. *Journal of Finance*, **55**, 1569–1622.

Tan, W.-Y. (2002) *Stochastic models with applications to genetics, cancers, AIDS and other biomedical systems*, vol. 4 of *Series on Concrete and Applicable Mathematics*. River Edge, NJ: World Scientific Publishing Co. Inc.

Vasicek, O. (1977) An equilibrium characterization of the term structure. *Journal of Financial Economics*, **5**, 177–188.

Wilkinson, D. J. (2006) *Stochastic modelling for systems biology.* Chapman & Hall/CRC Mathematical and Computational Biology Series. Chapman & Hall/CRC, Boca Raton, FL.