A methodological framework for Monte Carlo probabilistic inference for diffusion processes

Omiros Papaspiliopoulos*

June 26, 2009

1 Summary

The methodological framework developed and reviewed in this article concerns the *unbiased* Monte Carlo estimation of the transition density of a diffusion process, and the *exact* simulation of diffusion processes. The former relates to auxiliary variable methods, and it builds on a rich generic Monte Carlo machinery of unbiased estimation and simulation of infinite series expansions which relates to techniques used in diverse scientific areas such as population genetics and operational research. The latter is a recent significant advance in the numerics for diffusions, it is based on the so-called Wiener-Poisson factorization of the diffusion measure, and it has interesting connections to exact simulation of killing times for the Brownian motion and interacting particle systems, which are uncovered in this article. A concrete application to probabilistic inference for diffusion processes is presented by considering the continuous-discrete non-linear filtering problem.

2 Introduction

We consider statistical inference for models specified by stochastic differential equations (SDEs). SDEs provide a natural model for processes which (at least conceptually) evolve continuously in time and have continuous sample paths. From a more pragmatic point of view they offer a flexible framework for modelling irregularly spaced time series data. As a result they are used extensively throughout science; some examples (where statistical analysis of SDEs is considered) include finance (Sundaresan, 2000; Eraker et al., 2003; Ait-Sahalia and Kimmel, 2007), biology (Golightly and Wilkinson, 2006), molecular kinetics (Horenko and Schütte, 2008; Kou et al., 2005), and they are increasingly used in more mainstream statistical applications, e.g longitudinal data analysis (Taylor et al., 1994), space-time models (Brown et al., 2000) and functional data analysis (see for example Ramsay et al., 2007, and discussion therein). Specifically, an SDE for a d-dimensional process $V \in R^d$ is specified as follows:

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

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*. Boundary conditions are needed to complete the model specification. Certain assumptions are required on b and σ to ensure that (1) has a unique weak solution, see for example

*Department of Economics, Universitat Pompeu Fabra, omiros.papaspiliopoulos@upf.edu

CRiSM Paper No. 09-31, www.warwick.ac.uk/go/crism

Theorem 5.2.1 of Øksendal (1998) (although typically the globally Lipschitz and linear growth bounds are stronger than necessary). The unique solution is known as a diffusion process. It can be shown that it is a strong Markov process, thus it shares the Markov semigroup property with the solutions of ordinary differential equations (ODEs), which are obtained in the no-noise limit $\sigma = 0$. Note that the dimension of the driving Brownian motion can differ from that of the state process; in statistical applications an interesting possibility is to take $d > m$, in which case certain functions of V have smooth (at least once differentiable) sample paths. For example, we can model a process with smooth sample paths by specifying an SDE on the process and its time-derivatives. This gives rise to the so-called hypo-elliptic diffusion processes (Pokern et al., 2009). A simple popular hypo-elliptic model is the integrated Brownian motion which is often used in target tracking applications (see for example Gordon et al., 1993) and it relates to inference for unknown regression functions (Wahba, 1983).

The diffusion process can be used to model directly observed data, or it can be used to model latent processes which relate to the observable via likelihood functions. Statistical inference in such contexts consists of estimating unknown parameters involved in the specification of the drift and the diffusion coefficient, and estimating the process itself when it is unobserved. We are interested in likelihood-based inference for the unknown parameters, i.e maximum likelihood and Bayesian methods, and probabilistic inference for the unobserved processes, i.e inference according to the conditional law of the process given the observed data, where the prior law is given by the SDE specification (1). To simplify the presentation we will refer to such estimation procedures as *probabilistic inference*.

A major difficulty with probabilistic inference for diffusion processes is the intractability of their dynamics at scales other than infinitesimally small. This is due to the fact that in very few cases the equation can actually be solved. At infinitely small time increments the process has Gaussian dynamics, but due to the non-linearity of the drift and the non-constant diffusion the dynamics at arbitrary time increments, which are effectively a non-linear convolution of Gaussian distributions, are intractable. Therefore, to put it more formally, the transition density of a diffusion process which is defined as

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

is typically intractable. This raises a major difficulty especially since in many applications the observed data are at frequencies where the Gaussian assumption for (2) is not at all satisfactory. For example it is known that when V is directly observed at a given frequency, quasi-maximum likelihood estimators of parameters based on the Gaussian assumption are in general inconsistent as the number of data goes to infinity (Florens-Zmirou, 1989). On the other hand, this difficulty has motivated exciting research for analytic and Monte Carlo approximations of (2), see for example Aït-Sahalia (2006), Hurn et al. (2007), Papaspiliopoulos and Roberts (2009) and references therein. Typically, these approaches involve systematic bias due to time and/or space discretizations.

The methodological framework developed and reviewed in this article concerns the *unbiased* Monte Carlo estimation of the transition density, and the *exact* simulation of diffusion processes. The former relates to auxiliary variable methods, and it builds on a rich generic Monte Carlo machinery of unbiased estimation and simulation of infinite series expansions. This machinery is employed in diverse application areas such as population genetics and operational research. The latter is a recent significant advance in the numerics for diffusions and it is based on the so-called

Wiener-Poisson factorization of the diffusion measure. It has interesting connections to exact simulation of killing times for the Brownian motion and interacting particle systems, which are uncovered in this article.

The methodological framework we develop leads to unbiased probabilistic inference for diffusion processes. Our focus is more on the methodology than on its specific application to inference. Nevertheless, for clarity we consider a specific estimation problem, that of filtering an unobserved non-linear diffusion process on the basis of noisy observations at discrete time-points (the so-called continuous-discrete non-linear filtering problem, see for example Del Moral and Miclo, 2000). An overview of how to combine this framework with standard computational algorithms such as the Expectation-Maximization (EM) and Markov chain Monte Carlo (MCMC) to perform likelihood-based inference for diffusions is given in Beskos et al. (2006b).

The rest of the article is organised as follows. Section 3 introduces the continuous-discrete non-linear filtering problem, which serves as a motivating example. The section introduces the idea of replacing unknown densities by positive unbiased estimators and its interpretation as an auxiliary variable technique. Section 4 gives a representation of the transition density for a class of diffusion processes, which is key to our framework. Section 5 shows how to use this representation to achieve exact simulation of diffusion processes. Section 6 provides further insights on the exact simulation by linking it to the simulation of killing times of Brownian motion exploiting the connection between the exponential distribution and the Poisson process. It also relates the construction to interacting particle systems. Section 7 gives a detailed account of the machinery involved in deriving unbiased estimators of the diffusion transition density. This machinery is interesting outside the context of SDEs and links to the literature are provided. Section 8 closes with a discussion on extensions and directions for future research.

3 Random weight continuous-discrete particle filtering

The development in this section follows to some extent Fearnhead et al. (2008). We consider that (1) is unobserved, but partial information is available at discrete times $0 < t_1 < t_2 < \cdots < t_n$ in terms of observations $y_1, y_2, \ldots, y_n$ which are linked to the diffusion via a likelihood function, $f(y_i|V_{t_i})$. We also elicit a prior distribution on the diffusion initial state, say $p_0(V_0)$. Hence, we have a continuously evolving *signal* modelled as a diffusion process, and discrete-time observations. We are interested in the recursive calculation of the so-called filtering distributions, i.e the sequence of posterior distributions $p(V_{t_i}|y_{1:i})$ which will be denoted by $\pi_i(V_{t_i})$, where by standard convention $y_{1:i} = (y_1, \ldots, y_i)$. This is known as as the continuous-discrete filtering problem, see for example Del Moral and Miclo (2000). To simplify notation in this section we will subscribe the discrete skeleton of V by i rather than t_i, i.e $V_i := V_{t_i}$. Hence, we actually deal with a discrete-time Markov chain V_i, $i = 0, \ldots, n$, observed with noise. Hence the problem of interest can be coined as a discrete-time filtering problem, as follows.

Using marginalization, the Bayes theorem and the Markov property we obtain the following fundamental filtering recursion:

$$\pi_{i+1}(V_{i+1}) \propto \int f(y_{i+1}|V_{i+1}) p_{t_i,t_{i+1}}(V_i, V_{i+1}) \pi_i(V_i) \mathrm{d}V_i \,. \tag{3}$$

The filtering recursions can be solved analytically (in the sense of being able to describe the distributions in terms of a finite number of parameters which can be

recursively computed) only in very specific cases. This is for example the case when V is the solution of a linear SDE (e.g the integrated Brownian motion mentioned in Section 2) and the observed data are the unobserved diffusion plus Gaussian error, in which case we can use the Kalman filter.

For non-linear models, however, the state-of-the-art is to approximate the filtering distributions using Monte Carlo. This approach, which is known as *particle filtering* (Doucet et al., 2001), is characterised by two main steps. First, an approximation of π_i by a discrete distribution, denoted by π_i^N, whose support is a set of N particles, $\{V_i^{(j)}\}_{j=1}^N$, with associated (un-normalized probability) weights $\{w_i^{(j)}\}_{j=1}^N$. Substituting π_i^N for π_i in (3), yields a (continuous density) approximation to π_{i+1},

$$\tilde{\pi}_{i+1}(V_{i+1}) \propto \sum_{j=1}^{N} w_i^{(j)} f(y_{i+1}|V_{i+1}) p_{t_i,t_{i+1}}(V_i^{(j)}, V_{i+1})\,. \tag{4}$$

The aim of one iteration of the particle filter algorithm is to construct a further particle (discrete distribution) approximation to $\tilde{\pi}_{i+1}$. The second main step of the particle filter is to use importance sampling to sample from (4), thus obtain a particle approximation for $\tilde{\pi}_{i+1}$. A general framework for achieving this is given by the auxiliary particle filter of Pitt and Shephard (1999). We choose a proposal density of the form

$$\sum_{j=1}^{N} \beta_i^{(j)} q_{i+1}(V_{i+1}|V_i^{(j)}, y_{i+1})\,, \tag{5}$$

where the β_is are probabilities, and the q_is probability density functions. The algorithm is given in Table 1. Step PF2 of the algorithm includes a decision to *resample*

PF0 Simulate $V_0^{(j)} \sim p_0(V_0)$, and set $w_0^{(j)} = 1/N$, for $j = 1, \ldots, N$.
For $i = 0, \ldots, n-1$, for $j = 1, \ldots, N$:

PF1 calculate the effective sample size of the $\{\beta_i^{(k)}\}$,
$ESS = (\sum_{k=1}^N (\beta_i^{(k)})^2)^{-1}$; if $ESS < C$, for some fixed constant C,
simulate k_{i+1}^j from $p(k) = \beta_i^{(k)}$, $k = 1, \ldots, N$ and set $\delta_{i+1}^{(j)} = 1$;
otherwise set $k_{i+1}^j = j$ and $\delta_{i+1}^{(j)} = \beta_i^{(j)}$;

PF2 simulate $V_{i+1}^{(j)}$ from $q_{i+1}(\cdot \mid V_i^{k_{i+1}^j}, y_{i+1})$;

PF3 assign particle $V_{i+1}^{(j)}$ a weight

$$w_{i+1}^{(j)} = w_i^{(k_{i+1}^j)} \frac{\delta_{i+1}^{(j)} f(y_{i+1}|V_{i+1}^{(j)})\, p_{t_i,t_{i+1}}(V_i^{(k_{i+1}^j)}, V_{i+1}^{(j)})}{\beta_i^{(k_{i+1}^j)}\, q_{i+1}(V_{i+1}^{(j)} \mid V_i^{(k_{i+1}^j)}, y_{i+1})}\,. \tag{6}$$

Table 1: Auxiliary particle filter for state-space models

among existing particles when the variance of the proposal weights β exceeds a certain threshold. The decision is taken using the effective sample size, see for example Chapter 2 of Liu (2008). Note that taking $C < N$ and $\beta_i^{(k)} = 1/N$, resampling is never performed and the approach reduces to a direct importance sampling with

target π_{i+1} and proposals generated *independently* from $\prod_{k=0}^{i+1} q_k$. The (at least occasional) resampling, however, which introduces *dependence* among the particles, is crucial to break the curse of dimensionality inherent in an importance sampling algorithm. The resulting particle filter has good theoretical properties including consistency (Crisan, 2001) and central limit theorems for estimates of posterior moments (Del Moral and Miclo, 2000; Chopin, 2004; Künsch, 2005), as $N \to \infty$. Under conditions relating to exponential forgetting of initial conditions for the signal, particle filter errors stabilise as $n \to \infty$ (Del Moral and Guionnet, 2001; Künsch, 2005). Additionally, the filtering distributions are obtained at computational cost $\mathcal{O}(N)$, and unbiased estimators of the normalising constants (important in parameter estimation and model comparisons) are readily available. Improvements on independent sampling in PF1 can be made: see inter alia the stratified sampling ideas of Carpenter et al. (1999).

The algorithm in Table 1 applies generally to state-space time-series models. However, when the signal is a discretely-sampled diffusion process, the particle filter cannot be applied due to the intractability of the system transition density, which is necessary in the calculation of the weights. One way to by-pass this problem is to simulate the particles V_{i+1} according to the diffusion dynamics; then the transition density cancels out from (6). This requires the *exact simulation* of diffusions, which is discussed in Section 5. Another possibility is to try to obtain *unbiased estimators* for the transition density $p_{s,t}(u,v)$ for arbitrary s,t,u,v. The unbiasedness is needed to ensure that the particles are *properly weighted* (see for example Section 2.5.4 of Liu, 2008).

Section 7 shows how for each pair (u,v) and times (s,t), with $s<t$, to simulate *auxiliary* variables Ψ according to a distribution $Q(\,\cdot\,;s,t,u,v)$, and specify a computable function $r(\Psi,s,t,u,v)$, with the property that $\mathsf{E}(r(\Psi,s,t,u,v)) = p_{s,t}(u,v)$. Then, the so-called *random weight particle filter* (RWPF) inserts a further step between PF2 and PF3: simulate $\Psi_{i+1}^{(j)}$ from $Q(\,\cdot\,;t_i,t_{i+1},V_i^{(k_{i+1}^j)},V_{i+1}^{(j)})$ and compute $r(\Psi_{i+1}^{(j)},t_i,t_{i+1},V_i^{(k_{i+1}^j)},V_{i+1}^{(j)})$. This quantity replaces the intractable transition density in (6).

When r is positive this formulation has an interpretation as an expansion of the state-space using auxiliary variables. According to our construction, conditionally on V_i and V_{i+1}, Ψ_{i+1} is independent of Ψ_j and V_j for any j different from $i, i+1$. Additionally, it follows easily from the unbiasedness and positivity of r that for any u, $r(\psi,t_i,t_{i+1},u,v)$ is a probability density function as a function of (ψ,v) with respect to the product measure $Leb(\mathrm{d}v)\times Q(\mathrm{d}\psi;,t_i,t_{i+1},u,v)$, where Leb denotes the Lebesgue measure. Consider now an alternative discrete-time Markov model with unobserved states $(V_i,\Psi_i), i=1,\ldots,n$, transition density r and observed data y_i with observation density $f(y_i \mid V_i)$. By construction the marginal filtering distributions of V_i in this model are precisely π_i. Consider an auxiliary particle filter applied to this model where we choose with probabilities $\beta_i^{(j)}$ each of the existing particles $(V_i^{(j)},\Psi_i^{(j)})$, and generate new particles in the following way: V_{i+1} is proposed from q_{i+1} as described before, and conditionally on this value, Ψ_{i+1} is simulated according to Q. Then, it can be checked that the weight assigned to each particle is precisely that in the RWPF. Therefore, the RWPF is equivalent to an auxiliary particle filter on this discrete-time model whose latent structure has been augmented with the auxiliary variables Ψ_i. It is worth mentioning that the potential of using unbiased estimators of intractable densities while retaining the "exactness" of Monte Carlo algorithms is being increasingly recognised. The idea already appears in a disguised form in the auxiliary particle filter of Pitt and

Shephard (1999) and explicitly in the rejection control algorithm (see for example Section 2.6.1 of Liu, 2008). Beskos et al. (2006b) elaborate on this idea to design approximation-free MCMC algorithms for probabilistic inference for diffusions, Møller et al. (2006); Murray et al. (2006) devise novel MCMC algorithms for parameter estimation for models with intractable normalizing constants (which are functions of the parameters), Andrieu and Roberts (2009) develop and analyze theoretically a general class of MCMC algorithms where the target density is replaced by an importance sampling estimator, and Andrieu et al. (2008) show how to obtain exact MCMC algorithms for state-space models when the likelihood is estimated by the particle filter. Additionally, Papaspiliopoulos and Sermaidis (2007) show that the Monte Carlo EM algorithm can be adjusted using these ideas to increase monotonically an objective function.

Clearly, the replacement of an importance sampling weight, say w, with an unbiased estimator, say r, increases the variance: $\mathsf{V}(r) = \mathsf{V}(w) + \mathsf{E}(\mathsf{V}(r \mid w))$, since $\mathsf{E}(r \mid w) = w$, provided all variances exists. The expression suggests that the random weight importance sampler will be most efficient when $\mathsf{E}(\mathsf{V}(r \mid w))$ is relatively small compared to $\mathsf{V}(w)$.

In the auxiliary particle filter formulation given in Table 1 the positivity of the estimators is not necessary, since the resampling probabilities are controlled by β_i. Therefore, even if the actual weights w_i are negative, the algorithm in principle can still be carried out and yield consistent estimates of expectations over the filtering distributions. Clearly, in this case the w_is lose their interpretation as un-normalized probabilities; this is further discussed in Section 8. On the other hand, the generic algorithm in Table 2, proposed originally in Fearnhead et al. (2007), can be applied to ensure the positivity of the unbiased estimators. Suppose that we have N particles with true but unknown weights $w^{(j)}$ and for each j, let $r^{(i,j)}$ $i = 1, 2, \ldots,$ be a sequence of conditionally independent unbiased estimators of $w^{(j)}$. The procedure yields a random weight $r^{(j)} = \sum_{i=1}^{\tau} r^{(i,j)}$, where τ is a

(W0) Set $i = 1$, simulate $r^{(1,j)}$ and set $r^{(j)} = r^{(1,j)}$, for all $j = 1, \ldots, N$.

(W1.1) If $\min_j\{r^{(j)}\} > 0$ then STOP;

(W1.2) $i := i + 1$, simulate $r^{(i,j)}$ and set $r^{(j)} = r^{(j)} + r^{(i,j)}$, for all $j = 1, \ldots, N$.

Table 2: Creating positive unbiased importance weights exploiting Wald's identity

stopping time which depends on the sign of all weights. If $\mathsf{E}(\tau) < \infty$, then $\mathsf{E}(r^{(j)} \mid w^{(j)}) = \mathsf{E}(\tau) w^{(j)}$; this follows from Wald's identity, see Theorem 2 of Fearnhead et al. (2007). The intractable normalizing constant $\mathsf{E}(\tau)$ in the weights creates no problems, since it is common to all particles and will be cancelled out when the particle weights are re-normalized.

4 Transition density representation for a class of diffusions

The exact simulation and unbiased estimation methods developed in the article critically rely on a representation of the diffusion transition density. The representation relies on certain assumptions. To simplify exposition, we will assume from now on that (1) is time-homogeneous.

(A1) In the SDE (1), $d = m$, $\sigma = I$, and b is of *gradient form*, i.e there exists a function $U : R^d \to R$ (known as the potential) such that $b = \nabla U$.

The assumptions in (A1) are easily satisfied when $d = 1$. In that case, the assumption on b reduces to a differentiability condition. Additionally, when $\sigma(v)$ is a differentiable function of v, V can be transformed to a process with unit diffusion coefficient, by applying the transformation $v \to x = \int^v 1/\sigma(u)\mathrm{d}u$. Therefore, (A1) is restrictive only in multi-dimensional settings. Hence, in the rest of the paper we will consider a d-dimensional diffusion process X which solves the following SDE:

$$\mathrm{d}X_s = \nabla U(X_s)\,\mathrm{d}s + \mathrm{d}B_s, \quad s \in [0, T]\,; \tag{7}$$

where B is a d-dimensional Brownian motion, and $X_0 = x$. In the sequel X will also be used to denote an arbitrary continuous path, its meaning will be clear from the context.

Let $\mathbb{P}_0$ denote the law of the Brownian motion on the space of continuous paths, and $\mathbb{P}_b$ denote the probability law of X implied by (7). We can appeal to the Cameron-Martin-Girsanov theorem for Itô processes (see for example Theorem 8.6.6 of Øksendal, 1998) to obtain the likelihood ratio between the two measures on the time increment $[0, t]$. Applying also integration by parts facilitated by the gradient form of the drift, we obtain

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}|_t(X) = \exp\left\{U(X_t) - U(x) - \int_0^t \phi(X_s)\mathrm{d}s\right\}, \tag{8}$$

where $\phi(u) := (||b(u)||^2 + \Delta U(u))/2$, Δ is the Laplacian operator and $||\cdot||$ the Euclidean norm. Let $\mathbb{P}^*_{b;t,y}$ and $\mathbb{P}^*_{0;t,y}$ denote the laws on $[0, t]$ of X and B respectively, conditioned two hit at time t the value $y \in R^d$. A diffusion process conditioned to start and finish at specific values is known as a *diffusion bridge*.

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 X_t:

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}|_t(X) = \frac{p_{0,t}(x, y)}{\mathcal{G}_{0,t}(x, y)} \frac{\mathrm{d}\mathbb{P}^*_{b;t,y}}{\mathrm{d}\mathbb{P}^*_{0;t,y}}(X)\,, \tag{9}$$

where $\mathcal{G}_{0,t}(x, y)$ is the Gaussian transition density of the dominating Brownian motion. Then, re-arranging have the fundamental identity which underpins the methodological framework we develop here:

$$\frac{\mathrm{d}\mathbb{P}^*_{b;t,y}}{\mathrm{d}\mathbb{P}^*_{0;t,y}}(X) = \frac{\mathcal{G}_{0,t}(x, y)}{p_{0,t}(x, y)} \exp\left\{U(y) - U(x) - \int_0^t \phi(X_s)\mathrm{d}s\right\}. \tag{10}$$

Re-arranging (10) and taking expectations on both sides with respect to $\mathbb{P}^*_{0;t,y}$, we obtain the following representation for the transition density:

$$p_{0,t}(x, y) = \mathcal{G}_{0,t}(x, y) \exp\{U(y) - U(x)\} \mathsf{E}_{\mathbb{P}^*_{0;t,y}}\left[\exp\left\{-\int_0^t \phi(X_s)\mathrm{d}s\right\}\right]. \tag{11}$$

Therefore, we obtain the transition density as an expectation of an exponential transformation of a path integral, where the expectation is taken over the *Brownian bridge* measure.

The derivation of the likelihood ratio for diffusion bridge measures (10) can be made formal, see for example Theorem 2 of Delyon and Hu (2006). On a more general level, (11) follows from the basic principles of conditional expectation. In

particular let $(\Omega, \mathcal{F})$ be a measurable space, $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on the space with Radon-Nikodym derivative $\xi = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Then, the derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ restricted to $\mathcal{G}$ is $\mathsf{E}[\xi \mid \mathcal{G}]$. This is a very classical result which can be used to establish the existence of conditional expectation. On the other hand, assuming the existence of conditional expectation (using the projection approach, see for example Williams, 1991), the result follows from the definition of conditional expectation and the tower property of iterated conditional expectations. This basic result is instrumental in the statistical analysis of partially observed stochastic processes; for example Dembo and Zeitouni (1986) use it (see Section 2 of their article) to define an EM algorithm for partially observed continuous-time Markov processes. To obtain (11) we specify $\mathcal{G}$ as the σ-algebra generated by X_t and use the result in conjunction with (8).

For a thorough presentation of Monte Carlo identities for transition densities of diffusions, treatment of the general time-inhomogeneous and multivariate case and the historical development of these results see Sections 3 and 4 of Papaspiliopoulos and Roberts (2009).

5 Exact simulation of diffusions

Beskos et al. (2006a) and Beskos et al. (2008) recognised that (8) suggests an algorithm for the exact simulation of diffusion sample paths using rejection sampling. The algorithm is known generally as the *Exact Algorithm (EA)* and appeared in the literature in three generations corresponding to successive relaxations on the conditions which it requires; EA1 and EA2 are presented in Beskos et al. (2006a) and EA3 in Beskos et al. (2008).

For the development of the EA two further (relatively mild) assumptions are needed.

(A2) The function ϕ in (11) is lower bounded; let $\ell := \inf_u \phi(u) > -\infty$.

(A3) The function $\rho(y) := \exp\{U(y) - ||y - x||^2/(2t)\}$ is integrable in y for some t and for all x.

To avoid unnecessary notation, let us redefine ϕ as

$$\phi(u) = (||b(u)||^2 + \Delta U(u))/2 - \ell \geq 0\,. \tag{12}$$

We fix a time horizon t, such that (A3) holds, and consider the problem of simulating X_t according to the solution of (7) given $X_0 = x$, or equivalently according to the transition distribution $p_{0,t}(x, y)\mathrm{d}y$. (A3) allows us to define the so-called biased Wiener measure on the space of continuous paths on $[0, t]$ by its Radon-Nikodym derivative with respect to $\mathbb{P}_0$,

$$\frac{\mathrm{d}\mathbb{Z}}{\mathrm{d}\mathbb{P}_0}|_t(X) = \exp\{U(X_t)\}\,, \tag{13}$$

that is $\mathbb{Z}$ is obtained from $\mathbb{P}_0$ by biasing the marginal distribution of the latter at time t using the potential function U. Conditionally on the end-point, the two measures are identical. Then, by piecing everything together we have that

$$\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{Z}}|_t(X) \propto \exp\left\{-\int_0^t \phi(X_s)\mathrm{d}s\right\} \leq 1\,. \tag{14}$$

Therefore, there exists a rejection sampling algorithm on the path space for simulating diffusion sample paths $(X_s\,, 0 \leq s \leq t)$ according to $\mathbb{P}_b$ using proposals

from $\mathbb{Z}$ and accepting them with probability (14). Nevertheless, it is far from obvious how to carry out such an algorithm on the computer, i.e using a finite number of steps. This can be achieved by benefiting from a seemingly remarkable connection between the Brownian motion and the Poisson process, contained in the following Theorem (Beskos et al., 2008).

Theorem 1 *(Wiener-Poisson factorization) Let $\mathbb{L}$ denote the law of a unit rate Poisson process on $[0,t] \times [0,\infty)$ and define the extended law $\mathbb{Z} \otimes \mathbb{L}$ with typical realisation (X, Φ), with $\Phi = \{(\chi_j, \psi_j)\}_{j \geq 1}$, and $\{\psi_j\}$ non-decreasing. Define the event,*

$$\Gamma := \bigcap_{j \geq 1} \{\phi(X_{\chi_j}) < \psi_j\} \; . \tag{15}$$

Then, $\mathbb{P}_b$ on $[0,t]$ is the marginal distribution of X when $(X, \Phi) \sim \mathbb{Z} \otimes \mathbb{L} \,|\, \Gamma$.

Effectively, the Theorem formalizes the observation that the exponential term in (14) can be identified as the probability that an independent Poisson process on $[0,t] \times [0,\infty)$ has no points under the epigraph of $s \to \phi(X_s)$ for a given path X. The connection between the Poisson process and the diffusion measure is investigated and motivated further in Section 6.

Given knowledge of the range of ϕ we can appeal to the principle of *retrospective sampling* (Papaspiliopoulos and Roberts, 2008) to provide an algorithm for the exact simulation of X_t which can be carried out using a finite amount of computation. Suppose for instance that ϕ is also upper bounded,

$$\text{there exists an } r < \infty \text{ such that } \sup_u \phi(u) < r \, . \tag{16}$$

Then, the conditioned posed by (15) is trivially satisfied by all points of the Poisson process with $\psi_j > r$, and only a finite number of comparisons have to be made to check the condition. Additionally, since Φ is independent of X, we can first simulate the Poisson process on $[0,t] \times [0,r]$ and unveil X at the times χ_j specified by the Poisson process. When (15) is satisfied the simulated skeleton of X (which contains X_t) is retained, otherwise it is rejected and the procedure is repeated. This amounts to generating pairs (X, Φ) according to $\mathbb{Z} \otimes \mathbb{L}$ and accepting them when $(X, \Phi) \in \Gamma$, where we have used the upper bound of ϕ and retrospective sampling to check the condition using finite computation. The algorithm is given in Table 3, and each accepted draw X_t is a sample from the target diffusion at time t. For convenience, the χ_js are now ordered whereas the ψ_js are not. Note also that Step 2 simulates from the finite-dimensional distributions of $\mathbb{Z}$.

When ϕ is unbounded, the joint simulation according to $\mathbb{Z}$ of X and a random box which contains it, is required. This is the EA3 which is described in detail in Beskos et al. (2008). The extra effort needed in EA3 comes at an increased computational cost: the careful and extensive numerical investigation in Peluchetti and Roberts (2008) suggests as a rule of thumb that EA3 is about 10 times slower than EA1. Since EA is based on rejection sampling, when applied directly to $[0,t]$ the computational effort necessary to yield a draw grows exponentially with t. However, this is not the true complexity of the algorithm. The Markov property permits an implementation of the algorithm which has $\mathcal{O}(t)$ complexity, since the time increment $[0,t]$ can be split and the EA be applied sequentially. A further interesting property is that the acceptance probability of the EA is roughly constant when applied to intervals t/d as d increases; this is a by-product of the gradient

1. Generate a Poisson process $0 < \chi_1 < \chi_2 < \cdots$ of rate r on $[0,t]$. Let κ be the number of points. Generate a sequence of uniform random variables $\psi_j \sim \mathrm{Uni}[0,r], j = 1,\ldots,\kappa$.

2. Simulate $X_t \sim \rho$ given in (A3). Simulate $\{X_{\chi_1},\ldots,X_{\chi_\kappa}\}$, according to the Brownian bridge started at $X_0 = x$ and finishing at X_t.

3. If $\psi_j > \phi(X_{\chi_j})$ for all $j \leq \kappa$ then accept X_t; otherwise return to 1.

Table 3: The Exact Algorithm for the simulation of X_t according to the SDE (7) when (16) holds (EA1)

structure of the drift and the form of (14). This argument is supported empirically in Peluchetti and Roberts (2008), who find that EA1 has complexity $\mathcal{O}(d)$ in the dimension of the target diffusion. On the other hand, the complexity of EA3 as a function of d is worse than linear due to maximizations needed in the implementation of the algorithm.

6 Exact simulation of killed Brownian motion

The Wiener-Poisson factorization in Theorem 1 appears at first striking since it connects the law of a diffusion process to that of the Brownian motion and an independent Poisson process. However, this result is less surprising given a representation of the class of diffusions (7) as killed Brownian motion; see for example Section 8.2 of Øksendal (1998) where also the connections to the Feynman-Kac formula are discussed. In particular, consider an exponentially distributed random variable $E \sim \mathrm{Exp}(1)$, independent of X and define the killing time T as the following function of E and X:

$$T = \inf\left\{ s : \int_0^s \phi(X_s)\mathrm{d}s = E \right\}, \tag{17}$$

where ϕ is given in (12). Thus,

$$\Pr\left[T > t \mid X\right] = \exp\left\{ -\int_0^t \phi(X_s)\mathrm{d}s \right\}. \tag{18}$$

Then, it is easy to see that the scheme described in Table 4 yields an importance sampling approximation of the law of X_t induced by the SDE (7). The resulting weighted sample $\{(X_t^{(j)}, w_t^{(j)})\}_{j=1}^N$ is a particle approximation of the law of X_t. The killing step (Step 3 in Table 4) ensures that the law of the path conditioned to be alive has a density with respect to the Wiener measure given by the right-hand-side of (18), and the weighting (Step 4) is necessary to ensure that the path has density proportional to (8). However, the scheme of Table 4 is not practically implementable, since it involves an infinite amount of simulation in Step 2.

Note that for a given X, T is the first arrival time of a time-inhomogeneous Poisson process with intensity $\phi(X_s)$. Assume now that (16) holds. Then, we can simulate T *exactly* by thinning a dominating Poisson process with intensity r. Let $0 < \chi_1 < \chi_2 < \cdots$, be the time-ordered arrival times of the dominating Poisson process. Then, if each arrival χ_j is accepted with probability $\phi(X_{\chi_j})/r$, T is the

Set $j = 0$. While $j < N$ repeat:

1. Generate $E \sim \text{Exp}(1)$;
2. Generate a Brownian path X started from x, and keep track of $\int_0^s \phi(X_s)\mathrm{d}s$. Stop when $s = t$
3. Rejection: If $\int_0^s \phi(X_s)\mathrm{d}s > E$ reject the path, goto 1;
4. Weighting: If $\int_0^s \phi(X_s)\mathrm{d}s < E$ then $j := j+1$, set $X_t^{(j)} = X_t$, $w_t^{(j)} = e^{U(X_t)}$. Goto 1.

Table 4: Importance sampling approximation of the law of X_t by killed Brownian motion

first accepted arrival time. The algorithm described in Table 5 is a modification of the one given in Table 4; we call it the Exact Killing (EK) algorithm. The resulting

Set $j = 0$. While $j < N$ repeat:

1. Set $\chi_0 = 0$, $i = 0$
2. Set $i := i + 1$, simulate χ_i.
3. Simulate X_{χ_i} given $X_{\chi_{i-1}}$ according to the Brownian motion dynamics. If $\chi_i > t$ then simulate X_t given X_{χ_i} and $X_{\chi_{i-1}}$ according to the Brownian bridge dynamics, set $j := j + 1$ and $X_t^{(j)} = X_t$, $w_t^{(j)} = e^{U(X_t)}$. Goto 1.
4. If $\chi_i < t$, simulate $\psi_i \sim \text{Uni}(0, r)$. If $\psi_i > \phi(X_{\chi_i})$, then goto 2, else goto 1.

Table 5: Exact Killing: Exact simulation of a killed Brownian motion using thinning

weighted sample $\{(X_t^{(j)}, w_t^{(j)})\}_{j=1}^N$ is again a particle approximation of the law of X_t obtained by rejection (killing) and weighting, but now the procedure can be carried out exactly using a finite number of uniform and Gaussian random variables. This is made feasible precisely by the thinning of a Poisson super-process with rate r and it relies on the assumption (16).

The algorithm in Table 5 has intriguing connections to other exact simulation schemes for Markov processes. For example, the thinning of a Poisson super-process is a main ingredient of the algorithm of Fearnhead and Sherlock (2006) for the exact simulation of discrete state-space continuous-time Markov chains conditioned to start and finish at specific states. Most relevant to this article, is its direct connection with EA1 given in Table 3. In fact, the two algorithms share exactly the same rejection step. EK needs to weight the accepted draws, whereas EA1 by fixing the final time t apriori, it includes this bias in the dynamics of the proposal process which are according to $\mathbb{Z}$.

On the other hand, EK gives a particle approximation to the flow of distributions $t \to \mathbb{P}_b|_t$. Since EK also relies on rejection sampling, the computational effort to yield a particle at time t increases exponentially with t. The Markov property can be exploited here as well, by defining time increments of size, δ say. If a particle is

alive at time $i\delta$ but dies before $(i+1)\delta$, a new path is restarted from the value it has at time $i\delta$ rather than re-starting from time 0. Provided that the variance of the weights w_t does not increase with t (note that they depend only on X_t rather than the whole history) the complexity of the algorithm is $\mathcal{O}(t)$.

One can avoid the rejections involved in EK at the expense of introducing dependence among the simulated particles. Let N be a population of particles which move freely according to the Brownian dynamics. To each particle j, we assign a death time T_j, as before. Once a particle dies, then a randomly chosen particle of the remaining ones duplicates and each branch evolves conditionally independently. Again, it is easy to see that we can construct a super-process with intensity $r \times N$, which will contain all possible death times of all particles. We simulate iteratively these arrivals, at each arrival time χ_i, we pick at random one of the existing particles, j say, and propose to kill it. To do that, we realize its value at that time, we simulate $\psi_i \sim \text{Uni}[0, r]$, and check if $\psi_i < \phi(X^{(j)}_{\chi_i})$. If this is so we kill it and duplicate a randomly chosen one among the rest of the particles. If not, the particle remains alive. It is clear from the lack of memory of the underlying super-process, that at each arrival time, and after checking for killing and possibly adjusting the population, we can forget everything that has happened and start again from the current population of particles. To obtain an importance sample approximation for $\mathbb{P}_b|_t$ we weight each alive particle $X^{(j)}_t$ time t with $w^{(j)}_t = e^{U(X_t)}$ weight. Hence, we can simulate exactly the genealogy of this interacting particle systems which tracks the law of the diffusion process.

The connections between EA and EK beyond (16) are under investigation.

7 Unbiased estimation of the transition density using series expansions

The machinery required for producing unbiased estimators of diffusion transition densities is very broad in its scope and it is only mildly linked to the structure of diffusion processes. The techniques we present here are intrinsically linked to the Monte Carlo solution to fixed point problems, see for example Griffiths and Tavare (1994) for applications in population genetics, Wagner (1989) in the context of solutions of Partial Differential Equations (PDEs), Doucet et al. (2008) for a recent contribution in the literature and references, and Section 2.5.6 of Liu (2008) for a gentle introduction to the idea. The purpose in this section is to develop all components separately, emphasizing their generic purpose, and then piece them all together to solve the problem of interest in this article. The decoupling of the techniques greatly simplifies the understanding of the final method but also suggests possibilities for improvements. The main components of the methodology can be identified as follows. i) Expansion of functions into power series. This allows the unbiased estimation of the function given unbiased estimators of its argument. The expansion of the exponential function and the so-called Poisson estimator are treated in Section 7.1. Some optimality issues for the estimator are discussed and biased alternatives mentioned. ii) Unbiased truncation of infinite series. There are various techniques for the unbiased estimation of an infinite sum, based either on importance sampling or on integration by parts (effectively application of Fubini's theorem) followed by importance sampling. This is treated in Section 7.2. iii) Further structure is available when the unbiased estimator of the exponential of a path integral of a Markov process is required. Compared to i) the added feature is the explicit dependence of the unbiased estimators of the argument of the function. This is explored in Section 7.3, which couples this material with ii) to yield

a general class of unbiased estimators. The richer structure allows a more insightful mathematical formulation of the problem, as one of importance sampling in a countable union of product spaces. This point of view leads to the fourth component of the methodology. iv) Simulation from certain probability measures defined on a countable union of product spaces. This is treated in Section 7.4, and provides the optimal importance sampling estimator for the problem posed in Section 7.3. This formalism links directly with the so-called Monte Carlo method for solving integral equation and fixed-point problems. This is outlined in Section 7.4. There, we argue that the power expansion idea and the technique for solving integral equation, although related, they are not equivalent. An illustration to the estimation of the transition density of the Cox-Ingersoll-Ross diffusion process, considering the unbiased estimator and various biased estimators, is presented in Section 7.6.

7.1 Power series expansions: the exponential function and the Poisson estimator

We consider two related problems. Let X be an unknown quantity, and let $\tilde{X}_j$ be independent (conditionally on X) unbiased estimators of X, i.e $\mathsf{E}(\tilde{X} \mid X) = X$ (we will use $\tilde{X}$ to denote a generic element of the sequence). We will also assume that the $\tilde{X}_j$s have a common finite absolute moment, $\mathsf{E}(|\tilde{X}| \mid X) < \infty$. In many examples the $\tilde{X}_j$s have the same distribution conditionally on X. Let f be a non-linear function. Then, we are interested in estimating a) $f(X)$ or b) $\mathsf{E}(f(X))$ when X is a random variable. In fact, we are typically interested in b), however the argument is the same for both cases, hence we consider the two problems jointly. When f is linear the problem is trivial. However, when f is a real analytic function there is still the possibility to get unbiased estimators via series expansions. We concentrate on the case where $f(x) = e^x$. Then, for any fixed c, we have

$$\begin{aligned} e^X &= e^c \sum_{i=0}^{\infty} (X-c)^i / i! \\ &= e^c \sum_{i=0}^{\infty} \mathsf{E}\left[\prod_{j=1}^{i} (\tilde{X}_i - c) \mid X \right] / i! \\ &= e^c \, \mathsf{E}\left[\sum_{i=0}^{\infty} \prod_{j=1}^{i} (\tilde{X}_i - c) / i! \right] \end{aligned} \tag{19}$$

where the product $\prod_{j=1}^{0}$ is defined to be equal to 1. The role of c will be discussed later. Note that the absolute moment assumption on the $\tilde{X}$ justifies the third step in the above argument by dominated convergence. Hence, the infinite sum is an unbiased estimator of e^X. Still, this is not a realizable estimator. The topic of truncating unbiasedly infinite sums becomes of pivotal importance and it is discussed in the following Section. At a more elementary level, one way to yield a feasible estimator is to recognize the similarity of the expression to an expectation of a Poisson random variable. In fact, it is easy to check directly that for any $\lambda > 0$,

$$e^{\lambda + c} \prod_{i=1}^{\kappa} \frac{\tilde{X}_j - c}{\lambda}, \quad \kappa \sim Po(\lambda) \tag{20}$$

is a realizable unbiased estimator of e^X. We term (20) the *Poisson estimator*. Its second moment is easy to work out provided that the $\tilde{X}_j$s have a common second

moment, $\mathsf{E}[\tilde{X}^2 \mid X] < \infty$:

$$\exp\left\{\lambda + 2c + \frac{1}{\lambda}\mathsf{E}\left[(\tilde{X} - c)^2 \mid X\right]\right\} \tag{21}$$

The two constants c and λ are user-specified and relate to the sign and the variance of the estimator. For example, if $\tilde{X}$ is lower bounded c can be chosen to make the Poisson estimator positive, if this is desired (see for example Section 3). However, with two degrees of freedom the question of optimality in terms of variance is ill-posed, as shown in the following Proposition whose proof is straightforward.

Proposition 1 *Optimal implementation of the Poisson estimator for estimating e^X: Taking $c = -\lambda$, and $\lambda \to \infty$, the variance of the estimator converges monotonically to 0 and the estimator converges to e^X in mean square sense.*

Working directly from (21) we have that for fixed c, the optimal choice for λ is $\mathsf{E}[(\tilde{X} - c)^2 \mid X]^{1/2}$, whereas for a given computational budget λ, c is optimally chosen as $X - \lambda$. These are not feasible estimators, but can guide good choices.

Note that a biased plug-in alternative estimator is available in this context, which is given by $\exp\{\sum_{j=1}^{N} \tilde{X}_j/N\}$, where N plays the role of λ in the Poisson estimator. Even in this simple context the comparison of the two estimators in mean square error is not obvious. We will see these two possibilities in the context of diffusions in Sections 7.3 and 7.6.

In most cases of interest X is a random variable and we are actually interested in estimating $\mathsf{E}(e^X)$ with respect to the law of X. The argument presented earlier can be repeated to show that (20) is unbiased for this quantity, however we need the stronger condition

$$\mathsf{E}(\exp\{\mathsf{E}(|\tilde{X}| \mid X)\}) < \infty \tag{22}$$

as a sufficient condition to justify the third step in the development. A sufficient condition to ensure a finite second moment for a given λ, c is

$$\mathsf{E}\left[\exp\left\{\frac{1}{\lambda}(\mathsf{E}(\tilde{X}^2 \mid X) - 2cX)\right\}\right] < \infty\,; \tag{23}$$

then, the expected value of (21) gives the second moment. In this case we need to average, M say, independent realizations of the estimator, hence the computational cost is on average λM and the choice of optimal allocation in terms of λ and M is non-trivial.

Furthermore, c and λ can be chosen to depend on X. Fearnhead et al. (2008) proposed such generalized Poisson estimators to ensure positivity of the estimators. The estimator and its variance have the forms specified above, the conditions however which ensure their existence have to be modified appropriately.

7.2 Unbiased truncation of infinite series

In the previous section an estimator was given in terms of an infinite sum in (19). To avoid the impossible computation we extracted an unbiased estimator of the sum by expressing it as an expectation of a Poisson random variable. It turns out that this is just one instance of a generic methodology for unbiased estimation of infinite sums. Abstracting, let us consider the problem of finding an unbiased estimator of

$$S = \sum_{k=1}^{\infty} \alpha_k\,, \tag{24}$$

where we assume that the sum is finite a.s. As in the previous section we might be interested in $\mathsf{E}(S)$ when the α_ks are random variables, but the argument follows in a similar way. There are (at least) three ways to obtain an unbiased estimator of (24), the two of which turn out to be equivalent.

Firstly, we can use importance sampling. Let $\beta_k > 0$ be probabilities, i.e. $\sum_k \beta_k = 1$. Then α_K/β_K is an unbiased estimator of S, where K is simulated according to $\Pr[K = k] = \beta_k$. If

$$S_a = \sum_{k=1}^{\infty} |\alpha_k| < \infty\,, \tag{25}$$

then Jensen's inequality shows that it is optimal to take $\beta_k = |\alpha_k|/S_a$.

An alternative argument to yield effectively the same estimator, but useful when using this machinery in more elaborate contexts (see for example Section 7.3), is to define a sequence of "killing" probabilities $0 < p_k < 1$, for $k = 1, 2 \ldots$. Then, consider a discrete-time survival process where death happens at each time k with probability p_k. Let K be the death time. Then,

$$\frac{\alpha_K}{\prod_{i=1}^{K-1}(1-p_i)p_K} \tag{26}$$

is an unbiased estimator of S. Note that $\Pr[K = k] = \prod_{i=1}^{k-1}(1-p_i)p_k$. It is easy to check that $\sum_k \prod_{i=1}^{k-1}(1-p_i)p_k \leq 1$; if the sum is strictly less than one then the $K = \infty$ has a positive probability, which then yields an infeasible estimator. If the sum is 1, then the two estimators we have discussed are equivalent and correspond to the representation of a distribution in terms of the probabilities or the hazard function. The importance sampling estimator is obtained by taking $\beta_k = \prod_{i=1}^{k-1}(1-p_i)p_k$. On the other hand, for given probabilities β_k, let G be the survival function, $G(k) = \sum_{i=k}^{\infty} \beta_i$. Then, taking $p_k = 1 - G(k)/G(k-1)$ yields the second estimator.

The third estimator is based on an application of Fubini's theorem, which can be applied in this context under (25). Let again β_k be probabilities with survival function G. Then,

$$\begin{aligned}\sum_k \alpha_k &= \sum_k \frac{\alpha_k}{G(k)} G(k) = \sum_k \frac{\alpha_k}{G(k)} \sum_{i=k}^{\infty} \beta_i \\ &= \sum_{k=1}^{\infty}\sum_{i=k}^{\infty} \frac{\alpha_k}{G(k)}\beta_i = \sum_{i=1}^{\infty}\sum_{k=1}^{i} \frac{\alpha_k}{G(k)}\beta_i\end{aligned}$$

which suggests the following unbiased estimator of S:

$$\sum_{k=i}^{K} \alpha_i/G(i) = \sum_{i=1}^{K} \alpha_i \Big/ \prod_{i=1}^{K-1}(1-p_i) \tag{27}$$

where K is simulated according to $\Pr[K = k] = \beta_k$, and the equality follows from the equivalent representation in terms of killing probabilities.

It should be clear that the Poisson estimator (20) corresponds to a very specific setting where we use the importance sampling estimator with Poisson proposal probabilities for estimating the infinite expansion. It should also be clear that the other schemes we have discussed in this Section can be used to provide unbiased estimators of e^X and its expected value. These alternative estimators start with (19) and apply a technique for the unbiased estimation of the infinite sum.

7.3 Unbiased estimation of the expected value of exponential functions of Markov process path integrals

A very interesting instance of the generic context of Section 7.1 is when X is a path integral of a Markov process. With a slight abuse of notation, suppose that we are interested in estimating

$$I(x,t) := \mathsf{E}\left[\exp\left\{\int_t^1 g(s,X_s)\mathrm{d}s\right\}\right], \quad t \leq 1 \tag{28}$$

where X is a Markov process in R^d, with explicit transition density $p_{s,t}(x,y)$, such that $X_t = x$. The upper limit of the integration can be arbitrary, here it is taken to be 1 for notational simplicity. This problem was considered by Wagner (1988) who solved as we describe below. His approach combines the power expansions with the unbiased estimation of infinite series. Notice that the the estimation problem in (28) is raised when considering the estimation of the transition density for the class of diffusion processes considered in Section 4; see (11) where X is the Brownian bridge. The use of the estimators for the estimation of diffusion transition densities was considered in Beskos et al. (2006b), see also Section 7.6.

By the standard Monte Carlo integration trick, we have that $(1-t)g(\chi, X_\chi)$ where $\chi \sim \mathrm{Uni}(t,1)$, is conditionally on X an unbiased estimator of the exponent in (28). Working precisely as in Section 7.1, under the sufficient condition

$$I_a(x,t) := \mathsf{E}\left[\exp\left\{\int_t^1 |g(s,X_s)|\mathrm{d}s\right\}\right] < \infty, \quad \text{for all } t \leq 1 \tag{29}$$

we get the following infinite-series unbiased estimator of (28):

$$\sum_{k=0}^{\infty}\int_t^1\cdots\int_{u_{n-1}}^1\int_{R^d}\cdots\int_{R^d}\prod_{i=1}^{n} g(u_i,x_i)p_{u_{i-1},u_i}(x_{i-1},x_i)\mathrm{d}x_n\cdots\mathrm{d}x_1\mathrm{d}u_n\cdots\mathrm{d}u_1\,, \tag{30}$$

with the convention that $x_0 = x, u_0 = t$. This infinite expansion can be treated with the machinery of Section 7.2 to yield feasible unbiased estimators of (28). For example, an importance sampling estimator based on $Po(\lambda(1-t))$ probabilities and simulation of X according to its transition density, yields the Poisson estimator

$$e^{(\lambda+c)(1-t)}\prod_{j=1}^{\kappa}\frac{g(\chi_j,X_{\chi_j})-c}{\lambda}, \quad \kappa \sim Po(\lambda(1-t)), \chi_j \sim \mathrm{Uni}(t,1)\,. \tag{31}$$

Note however that with the same variables we can consider the alternative estimator based on the application of Fubini's theorem discussed in Section 7.2, or indeed use a different proposal distribution for the index K (e.g the negative binomial).

The specific structure of the exponent in (28) (as opposed to the generic one in Section 7.1) permits a mathematically richer formulation of the estimation problem. This is done in Wagner (1988) (see in particular Propositions 1,2 and 4 of the article). This formulation casts the estimation of (28) as a familiar problem in Monte Carlo. Specifically, let us define the following union of product spaces, $\mathcal{Y} := \bigcup_{k=0}^{\infty}\mathcal{Y}_k$ where $\mathcal{Y}_k = \{k\}\times\mathcal{X}^{k+1}$, and in our context $\mathcal{X}$ is the space $[t,1]\times R^d$. Let us now define the following signed measure φ on $\mathcal{Y}$ indexed by (x,t), and given by the formulas

$$\begin{aligned}&\varphi(k,\mathrm{d}(t_0,x_0)\times\cdots\times\mathrm{d}(t_k,x_k)\,;x,t) = \delta_{(t,x)}(\mathrm{d}(t_0,x_0))\\ &\prod_{i=1}^{k}1_{t_i}[t_{i-1},1]g(t_i,x_i)p_{t_{i-1},t_i}(x_{i-1},x_i)\mathrm{d}t_i\mathrm{d}x_i\end{aligned} \tag{32}$$

where δ denotes the Dirac delta function, and $1_x[A]$ is 1 if $x \in A$ and 0 otherwise. In this formulation, (30) shows that (28) is the normalising constant of φ: $I(x,t) = \varphi(\mathcal{Y}; x, t)$, hence can reformulate the orginal problem as one of estimating a normalising constant. Importance sampling is one possibility to do this by constructing measures on $\mathcal{Y}$ and computing the Radon-Nikodym derivative between the two measures for the generated samples. Provided that the normalising constant of the proposal distribution is known, the weight assigned to each generated sample is an unbiased estimator of $\varphi(\mathcal{Y}; x, t)$. Summarising, the expansion in a power series and the explicit structure of the exponent allow the re-formulation of estimation of (28) as the computation of a normalising constant of a signed measure. The material of Section 7.2 together with standard Monte Carlo tecniques effectively it gives methods for constructing proposal distributions on $\mathcal{Y}$ to be used in the importance sampling. Wagner (1988) gives the following generic estimator where let $p_0(s,x) > 0$ is a killing probability and $q_{s,t}(x,y)$ is an alternative tractable transition density:

$$\prod_{i=1}^{K} \frac{g(\chi_i, X_{\chi_i}) p_{\chi_{i-1},\chi_i}(X_{\chi_{i-1}}, X_{\chi_i})}{(1 - p_0(\chi_{i-1}, X_{\chi_{i-1}})) q_{\chi_{i-1},\chi_i}(X_{\chi_{i-1}}, X_{\chi_i})} \frac{1}{p_0(\chi_K, X_{\chi_K})}, \tag{33}$$

where the χ_is are ordered uniforms on $[t,1]$ and the X_{χ_i} are generated according to the transitions q. Let $|\varphi$ be the total variation of φ, thus it is obtained by replacing g with its absolute value in the definition given above. Then, by Jensen's inequality (as in Section 7.2) it follows that the optimal proposal distribution in terms of minimising the variance of the estimator, is $|\varphi|/I_a(x,t)$. Simulation from probability measures in $\mathcal{Y}$ is treated in the next section.

We close the remark that alternative biased plug-in estimators (as discussed in Section 7.1) are available. For example

$$\exp\left\{ \frac{1-t}{N} \sum_{j=1}^{N} g(\chi_j, X_{\chi_j}) \right\} \tag{34}$$

with the random elements as in (31); alternative numerical approximation of the integral in the exponents can be considered. A comparison among different schemes is carried out in Section 7.6.

7.4 Simulation from probability measures on unions of spaces

The fourth main ingredient of the methodological framework for unbiased estimation is linked with the simulation from the following series of measures. Consider the following abstract problem: let $\gamma(x)$ be a positive function on $\mathcal{X}$; $p(x,y)$ be a transition density (i.e probability density in y and measurable in x), where $x, y \in \mathcal{X}$; and $\delta_x(dy)$ be the Dirac measure centred at x. Consider the product the space $\mathcal{Y} := \bigcup_{k=0}^{\infty} \{k\} \times \mathcal{X}^{k+1}$ with typical element $(k, x_0, x_1, \ldots, x_k)$ with the convention $x_0 = x$. We have already seen this context in Section 7.3, where $\mathcal{X} = [t,1] \times R^d$.

We define the following positive measure on $\mathcal{Y}$ indexed by $x \in \mathcal{X}$:

$$\nu(k, \mathrm{d}x_1 \times \cdots \times \mathrm{d}x_{k+1}\,; x) := \delta_x(\mathrm{d}x_0) \prod_{i=1}^{k} p(x_{i-1}, x_i) \gamma(x_i) \mathrm{d}x_1 \cdots \mathrm{d}x_{k+1}\,. \tag{35}$$

We assume that $I(x) := \nu(\mathcal{Y}\,; x) < \infty$, and define $\tilde{\nu}(\cdot; x) = \nu(\cdot; x)/I(x)$ to be the corresponding probability measure on $\mathcal{Y}$. Note that by definition $I(x) > 1$. The aim of this section is to simulate draws from $\tilde{\nu}$ and to show that distributions of this form

provide the optimal importance sampling distributions in the context of Section 7.3. The construction is theoretical, since it will typically not be feasible to carry out the simulation. Nevertheless, it provides insights on the optimal implementation of the unbiased estimators we consider in this article.

To start with note the fundamental recursion implied by the definition of the measures and the normalizing constants:

$$I(x) = 1 + \int_{\mathcal{Y}} I(x_1)p(x, x_1)\gamma(x_1)\mathrm{d}x_1 \,. \tag{36}$$

Using the same argument that lead to (36) we can obtain the following marginal-conditional distributions under $\tilde{\nu}$: $\tilde{\nu}(k = 0; x) = 1/I(x)$, $\tilde{\nu}(\mathrm{d}x_1, k > 0; x) \propto p(x, x_1)\gamma(x_1)I(x_1)\mathrm{d}x_1$. In the same way we obtain the general expressions

$$\begin{aligned}
\tilde{\nu}(k > i-1, \mathrm{d}x_1, \ldots, \mathrm{d}x_i; x) &= I(x_i) \prod_{j=1}^{i} p(x_{j-1}, x_j)\gamma(x_j) \\
\tilde{\nu}(\mathrm{d}x_i \mid x, x_1, \ldots, x_i, k > i-1) &= p(x_{i-1}, x_i)\gamma(x_i)I(x_i)\mathrm{d}x_1/(I(x_{i-1}) - 1) \\
\tilde{\nu}(k = i \mid x, x_1, \ldots, x_i, k > i) &= 1/I(x_i) \,.
\end{aligned}$$

The last two equations give the necessary structure for the simulation from $\tilde{\nu}$ using a Markov chain, by sequentially at each stage i first simulating a new value x_i and then deciding on whether to stop the simulation. The procedure results with a string $(k, x_0, x_1, \ldots, x : k)$. The problem of simulation from probability measures on $\mathcal{Y}$ with structure as $\tilde{\nu}$ was recently considered in Doucet et al. (2008) using trans-dimensional MCMC; see also the article for further references. This problem, together with the corresponding task of estimating the normalising constant comes up in a large number of scientific contexts. This is due to the fact that it is intrinsically related to the numerical solution of fixed point problems. This is described in the following section.

7.5 Monte Carlo for integral equations

Suppose that we are interested in the solution of the following integral equation:

$$I(x) = h(x) + \int_{\mathcal{X}} p(x, y)I(y)\mathrm{d}y \tag{37}$$

where h is explicitly known for all x. This type of equations (and their discrete-valued counterparts) appear in a variety of problems. We have already seen an instance; $I(x, t)$ in (28) satisfies such an equation with $h = 1$. By successive substitution of I in the equation we obtain the infinite expansion

$$I(x) = h(x) + \sum_{k=1}^{\infty} \int_{\mathcal{X}^k} \prod_{i=1}^{k} p(x_{i-1}, x_i)h(x_k)\mathrm{d}x_1 \times \cdots \mathrm{d}x_k \tag{38}$$

with the convention $x_0 = x$. The analogy with the problems treated in Sections 7.2-7.4 is direct. This is the reason why the same machinery which is used in the solution of the fixed point problems becomes useful in the unbiased estimation of the diffusion transition density. Nevertheless, the power expansions discussed in Section 7.1 do not necessarily lead to a fixed point problem. However, the techniques of Section 7.2 still apply to yield unbiased estimators even in these cases.

7.6 Illustrating example: the CIR density

We close the Section with an illustration of the methodology on the estimation of the transition density of the so-called Cox-Ingersoll-Ross (CIR) diffusion (Cox et al., 1985). This is a one-dimensional diffusion with b and σ in (1) given by $-\rho(x-\mu)$, and $\sigma\sqrt{x}$ respectively, where $\rho > 0, \sigma > 0, \mu$ are parameters and $x \in R_+$. This diffusion is not in the form (7) but it can be transformed as described in Section 4. When the transformation is applied the transition density of the original process is linked by a change of variables to the one of the transition density of the unit-diffusion-coefficient process; see Beskos et al. (2006b). However, in this model when the process is transformed its measure is absolutely constinuous with respect to the law of the Brownian motion conditioned to remain positive, which is known as the *Bessel process*. Therefore, (11) holds but the expectation is taken with respect to the law of the Bessel bridge. In our numerical results it turns out that it does not really make a difference whether one works with the Brownian or the Bessel bridge. We consider four estimators. First, the unbiased estimator obtained by using the Poisson estimator (31) to estimate the expectation in (11) (using Bessel bridge dominating measure). Second, a biased estimator based on Riemmann approximation of the exponent in (11). This is in the spirit of the plug-in estimators discussed in Section 7.3 but where the times to evaluate the path are chosen deterministically. This estimator in the context of diffusions was considered in Nicolau (2002). Finally, we consider two estimators obtained using the discrete-time approach of Durham and Gallant (2002). We use their estimator on the original CIR process and on the transformed to unit-diffusion-coefficient process. The estimator of Durham and Gallant (2002) applied to the transformed process is closely related to the estimator of Nicolau (2002): the only difference is that the latter applies numerical integration to a Cameron-Martin-Girsanov formula with the stochastic integral eliminated using integration by parts, whereas the former applies numerical integration on the expression which contains the stochastic integral.

The transition density of the CIR is explicitly known, hence it can be used to assess the root mean square error of the estimators. Our simulation setup is as follows. We consider the parameter values used in the simulation study in Durham and Gallant (2002): $\rho = 0.06, \mu = 0.5, \sigma = 0.15$ and starting point for the diffusion $X_0 = 0.1$. We consider two final times, a small one $t = 1/252$ and a large one $t = 1/2$, and we estimate the transition density for three different ending points which correspond to the 10, 50 and 90 percent quantiles of the transition distribution. For the biased estimators we consider various values for N, the number of evaluations on a given path, $N = 2^i, i = 2, 3, \ldots, 8$. For the Poisson estimator we choose the average computational cost to be the same as that of the biased estiamators and we take $c = \lambda$. In each case we average M independent realizations of the estimator, where we take $M = N^2$ following the asymptotic result of Stramer and Yan (2007). For the estimation of the root mean square error of each estimator we average 120 independent replicates. Figures 1 and 2 contain the results of the simulation, where we plot the logarithm of the root mean square error against the logarithm of the number of evaluations per path. The study shows the variance reduction effectuated by the expression of the transition density in (11). Moreover, the unbiased estimator works very well in this setup. In this article we have pursued unbiasedness due to its connection with auxiliary variable methods. Nevertheless, the results show that the estimator has comparable or better performance than biased alternatives.

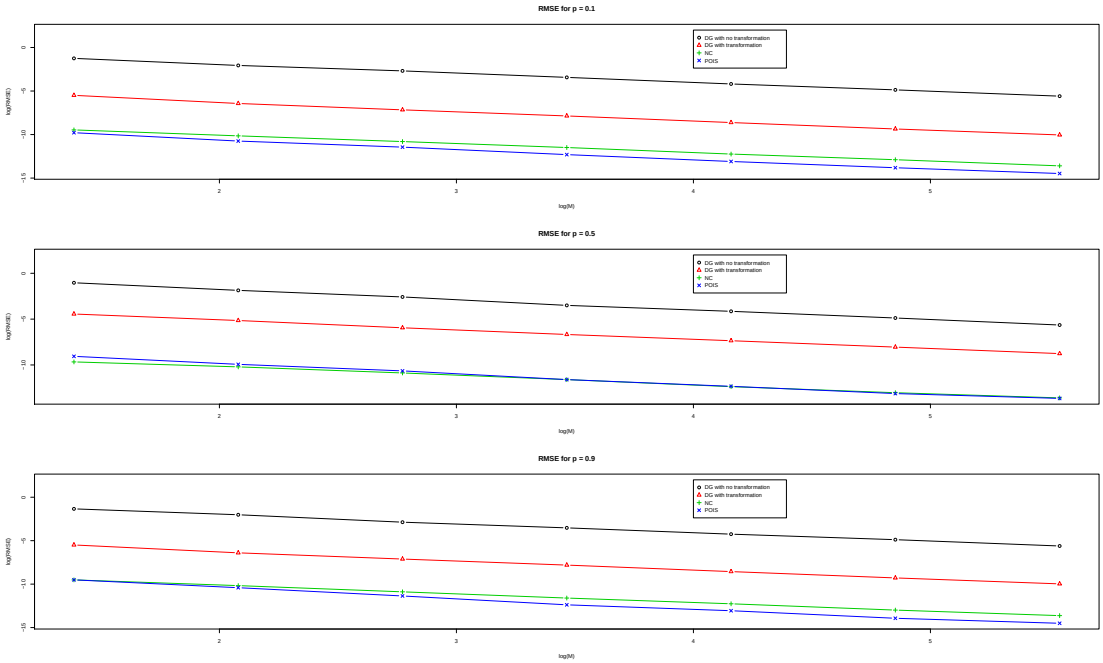

Figure 1: Logarithm of the root mean square error of the estimators against the logarithm of the number of imputed points per simulated path. The transition of the CIR process is estimated for three ending points corresponding to the 10 (top), 50 (middle) and 90 (bottom) quantiles of the transition distribution. The time increment is $t = 1/252$.

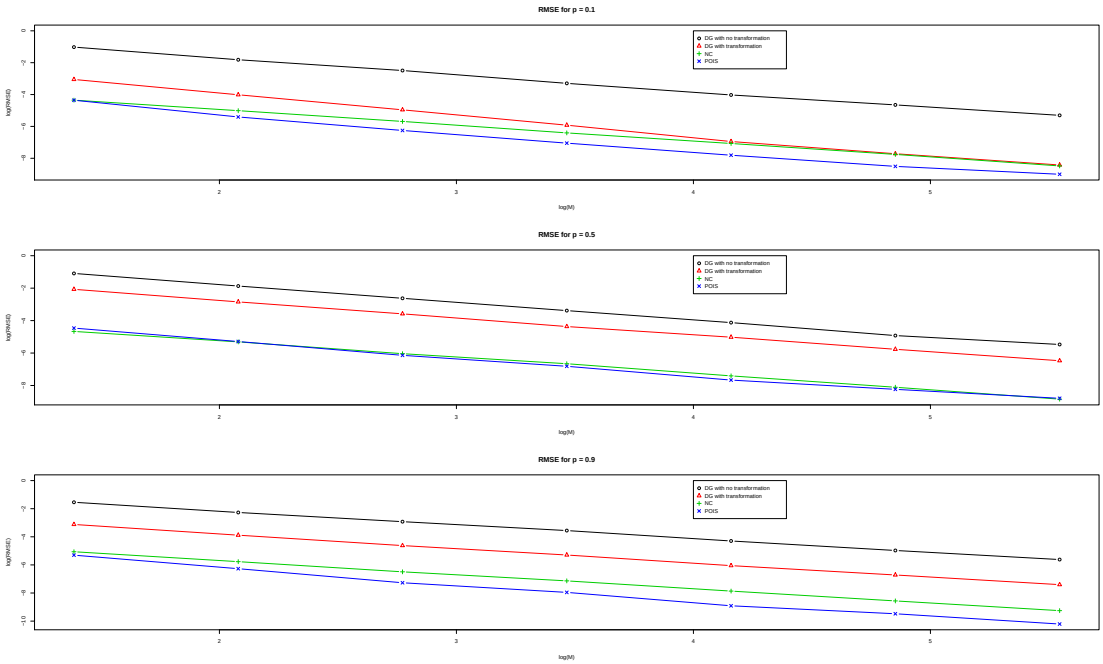

Figure 2: Details as in Figure 1 but $t = 1/2$.

8 Discussion and directions

We have reviewed and developed a rich methodological framework for the Monte Carlo assisted probabilistic inference for diffusion processes. On the one hand, the framework is based on representations of the diffusion process which can be exploited for its exact simulation. On the other hand, the framework relies on a generic importance sampling machinery which has been used in various other contexts. The Exact Algorithm and the Poisson estimator build bridges between these two aspects, see for example the discussion in Beskos et al. (2006b). It is interesting to understand deeper the connections; this might lead to new exact simulation algorithms outside the framework described in Section 5.

A further aspect of the methodology is the interplay between unbiased estimation of densities and the "exactness" of Monte Carlo schemes. We explored this possibility in Section 3 to derive a Sequential Monte Carlo algorithm for diffusions. The possibility to carry out such algorithms based on negative estimates of the weights is intriguing and little explored. The interpretation in terms of expansion of the state space with auxiliary variables is lost in this case. Such interpretation is in some cases instrumental in establishing the validity of Monte Carlo algorithms working with estimates of objective functions, see for example Andrieu and Roberts (2009).

Acknowledgements

The author would like to acknowledge financial support by the Spanish government through a "Ramon y Cajal" fellowship and the grant MTM2008-06660, the Berlin Mathematical School for hosting him as a visiting Professor while preparing this manuscript, Giorgos Sermaidis for useful suggestions and Christian Robert and Randal Douc for motivating discussions on the unbiased estimation techniques.

References

Aït-Sahalia, Y. (2006). Likelihood inference for diffusions: a survey. In *Frontiers in statistics*, pages 369–405. Imp. Coll. Press, London.

Ait-Sahalia, Y. and Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics*, 83(2):413–452.

Andrieu, C., Doucet, A., and Holenstein, R. (2008). Particle markov chain monte carlo. submitted.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient monte carlo computations. *Ann. Statist.*, 37(2):697–725.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2006a). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12:1077–1098.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2008). A new factorisation of diffusion measure and finite sample path constructions. *Methodol. Comput. Appl. Probab.*, 10(1):85–104.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006b). Exact and efficient likelihood–based inference for discretely observed diffusions (with Discussion). *J. Roy. Statist. Soc. Ser. B*, 68(3):333–82.

Brown, P. E., Kåresen, K. F., Roberts, G. O., and Tonellato, S. (2000). Blur-generated non-separable space-time models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(4):847–860.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). An improved particle filter for non-linear problems. *IEE proceedings-Radar, Sonar and Navigation*, 146:2–7.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32:2385–2411.

Cox, J. C., Ingersoll, Jr., J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407.

Crisan, D. (2001). Particle filters - a theoretical perspective. In Doucet, A., de Freitas, N., and gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, pages 17–41. Springer–Verlag; New York.

Del Moral, P. and Guionnet, A. (2001). On the stability of interactin processes with applications to filtering and genetic algorithms. *Ann. Inst. of H. Poincaré Probab. Statist.*, 37:155–194.

Del Moral, P. and Miclo, L. (2000). *Branching and interacting particle systems. Approximations of Feymann-Kac formulae with applicationc to non-linear filtering*, volume 1729.

Delyon, B. and Hu, Y. (2006). Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Process. Appl.*, 116(11):1660–1675.

Dembo, A. and Zeitouni, O. (1986). Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stochastic Process. Appl.*, 23(1):91–113.

Doucet, A., de Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 3–14. Springer, New York.

Doucet, A., Johansen, A., and Tadic, V. (2008). On solving integral equations using Markov Chain Monte Carlo. Available from http://www.cs.ubc.ca/ arnaud/TR.html.

Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *J. Bus. Econom. Statist.*, 20(3):297–338. With comments and a reply by the authors.

Eraker, B., Johannes, M., and Polson, N. (2003). The impact of jumps in volatility and returns. *Journal of Finance*, 58(3):1269–1300.

Fearnhead, P., Papaspiliopoulos, O., Roberts, G., and Stuart, A. (2007). Filtering systems of coupled stochastic differential equations partially observed at high frequency. under revision.

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.

Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov-modulated Poisson process. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(5):767–784.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557.

Golightly, A. and Wilkinson, D. (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13:838–851.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings Part F: Communications, Radar and Signal Processing*, 140:107–113.

Griffiths, R. and Tavare, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46:131–158.

Horenko, I. and Schütte, C. (2008). Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics. *Multiscale Model. Simul.*, 7(2):731–773.

Hurn, A., Jeisman, I., J., 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(3):390–455.

Kou, S. C., Xie, X. S., and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *J. Roy. Statist. Soc. Ser. C*, 54(3):469–506.

Künsch, H. R. (2005). Monte Carlo filters:Algorithms and theoretical analysis. *Annals of Statistics*, 33:1983–2021.

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

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Murray, I., Ghahramani, Z., and MacKay, D. (2006). MCMC for doubly-intractable distributions. *In Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-2006)*.

Nicolau, J. (2002). A new technique for simulating the likelihood of stochastic differential equations. *Econom. J.*, 5(1):91–103.

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

Papaspiliopoulos, O. and Roberts, G. (2009). Importance sampling techniques for estimation of diffusion models. In *SEMSTAT*. Chapman and Hall.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective markov chain monte carlo for dirichlet process hierarchical models. *Biometrika*, 95:169–186.

Papaspiliopoulos, O. and Sermaidis, G. (2007). Monotonicity properties of the monte carlo em algorithm and connections with simulated likelihood. Available from http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2007/paper07-24.

Peluchetti, S. and Roberts, G. (2008). An empirical study of the efficiency of ea for diffusion simulation. CRiSM Technical report 08-14, available from http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2008/paper08-14.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filters. *J. Amer. Statist. Assoc.*, 94(446):590–599.

Pokern, Y., Stuart, A. M., and Wiberg, P. (2009). Parameter estimation for partially observed hypoelliptic diffusions. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 71(1):49–73.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(5):741–796. With discussions and a reply by the authors.

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(4):483–496.

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

Taylor, J., Cumberland, W., and Sy, J. (1994). A Stochastic Model for Analysis of Longitudinal AIDS Data. *Journal of the American Statistical Association*, 89(427):727–736.

Wagner, W. (1988). Unbiased multi-step estimators for the Monte Carlo evaluation of certain functional integrals. *J. Comput. Phys.*, 79(2):336–352.

Wagner, W. (1989). Unbiased Monte Carlo estimators for functionals of weak solutions of stochastic differential equations. *Stochastics Stochastics Rep.*, 28(1):1–20.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B*, 45(1):133–150.

Williams, D. (1991). *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge.