Chapter 4

Monte Carlo Probabilistic Inference for Diffusion Processes: A Methodological Framework

Omiros Papaspiliopoulos

4.1 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 as a statistical model throughout science; for example in finance [48, 20, 2], biology [26], molecular kinetics [29, 31]. They are increasingly used in more mainstream statistical applications, e.g longitudinal data analysis [49], space-time models [8] and functional data analysis, see for example [45] and discussion therein. Specifically, an SDE for a \( d \)-dimensional process \( V \in \mathbb{R}^d \) is specified as follows:

\[
dV_s = b(s, V_s) \, ds + \sigma(s, V_s) \, dB_s, \quad s \in [0, T];
\]

(4.1)

\( B \) is an \( m \)-dimensional standard Brownian motion, \( b(\cdot, \cdot) : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d \) is called the drift, \( \sigma(\cdot, \cdot) : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d \times m} \) is called the diffusion coefficient. Boundary conditions are needed to complete the model specification. Certain assumptions are required on \( b \) and \( \sigma \) to ensure that Eq. (4.1) has a unique weak solution, see for example Theorem 5.2.1 of [38]. 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 \( d > m \). For example, we can model a process with differentiable sample paths by specifying an SDE on the process and its time-derivatives. This gives rise to the so-called hypo-elliptic diffusion processes [44]. A simple popular hypo-elliptic model is the integrated Brownian motion which is often used in target tracking applications, see for example [27], and it relates to inference for unknown regression functions [53].

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
Monte Carlo Probabilistic Inference for Diffusion Processes

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 (4.1). 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 the transition density:

\[ p_{s,t}(v, w) = \Pr \left[ V_t \in dw \mid V_s = v \right] / dw, \quad t > s, w, v \in \mathbb{R}^d. \]  

(4.2)

This is due to the fact that only in very few cases the SDE can be analytically solved. At infinitely small time increments, i.e for \( t-s \approx 0 \), Eq. (4.2) can be satisfactorily approximated by a Gaussian. However, this approximation is very poor for arbitrary time increments. Intuitively, the transition distribution for longer time increments is a non-linear convolution of Gaussian distributions, hence it is intractable. It is known that for fixed observation frequency, quasi-maximum likelihood estimators of parameters based on a first-order Gaussian approximation to Eq. (4.2) are in general inconsistent [24]. On the other hand, this difficulty has motivated exciting research for analytic and MC approximations of Eq. (4.2), see for example [1], [30], [40] 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 MC 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 MC 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 the so-called continuous-discrete non-linear filtering problem, see for example [14]. 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 [7].

The rest of the article is organised as follows. Section 4.2 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 4.3 gives a representation of the transition density for a class of diffusion processes, which is key to our framework. Section 4.4 shows how to use this representation to achieve exact simulation of diffusion processes. Section 4.5 provides further insights to 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 4.6 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 4.7 closes with a discussion.
4.2 Random Weight Continuous-Discrete Particle Filtering

The development in this section follows to some extent [21]. We consider that Eq. (4.1) 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 \(p(V_{t_1}|y_{1:t_1})\) which will be denoted by \(\pi(V_{t_i})\), where by standard convention \(y_{1:t} = (y_1, \ldots, y_t)\). This is known as as the continuous-discrete filtering problem, see for example [14]. 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:

\[
\pi_{i+1}(V_{i+1}) \propto \int f(y_{i+1}|V_{i+1})p_{V_i,V_{i+1}}(V_i, V_{i+1})\pi_i(V_i)dV_i. \tag{4.3}
\]

Only in very specific cases the filtering distributions can be characterized by a finite number of parameters which can be recursively computed. A model amenable to this type of analysis is obtained when \(V\) is the solution of a linear SDE (e.g. the integrated Brownian) observed with additive Gaussian error. In this case we can use the Kalman filter to do the computations.

For non-linear models, however, the state-of-the-art is to approximate the filtering distributions using MC. The corresponding algorithms, known as particle filters (PFs) [17], are characterised by two main steps. First, an approximation of \(\pi_i\) by a discrete distribution, denoted by \(\pi_i^N\), whose support is a set of \(N\) particles, \(\{V_i^{(j)}\}_{j=1}^N\), with associated (un-normalized probability) weights \(w_i^{(j)}\). Substituting \(\pi_i^N\) for \(\pi_i\) in Eq. (4.3), yields a (continuous density) approximation to \(\pi_{i+1}\),

\[
\tilde{\pi}_{i+1}(V_{i+1}) \propto \sum_{j=1}^N w_i^{(j)} f(y_{i+1}|V_{i+1})p_{V_i,V_{i+1}}(V_i^{(j)}, V_{i+1}) . \tag{4.4}
\]

The aim of one iteration of the PF is to construct a further particle (discrete distribution) approximation to \(\tilde{\pi}_{i+1}\). The second main step of the PF is to use importance sampling to sample from Eq. (4.4), thus obtain a particle approximation for \(\tilde{\pi}_{i+1}\). A general framework for achieving this is given by the auxiliary particle filter of [43]. We choose a proposal density of the form

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

where the \(\beta_i\)s are probabilities, and the \(q_i\)s probability density functions. The algorithm is given in Algorithm 4.1. Step PF2 of the algorithm includes a decision to resample among existing particles when the variance of the proposal weights \(\beta\) exceeds a certain threshold. The decision is taken using the effective sample size, see for example Chapter 2 of [33]. Note that taking \(C < N\) and \(\beta_i^{(k)} = 1/N\), resampling is never performed and the approach
Algorithm 4.1 Auxiliary PF for state space models.

PF0 Simulate \( V_i^{(j)} \sim p_0(V_0) \), and set \( w_i^{(j)} = 1/N \), for \( j = 1, \ldots, N \).

for \( i = 0, \ldots, n - 1 \), \( j = 1, \ldots, N \) do

[PF1] calculate the effective sample size of the \( \{ \beta_i^{(k)} \} \).

\[ \text{ESS} = \left( \sum_{k=1}^{N} (\beta_i^{(k)})^2 \right)^{-1} \; \text{if ESS} < C, \text{ for some fixed constant} C, \text{ simulate} \; k_i^{(j)} \text{ from} \; p(k) = \beta_i^{(k)}; \; k = 1, \ldots, N \; \text{and set} \; \delta_i^{(j)} = 1; \; \text{otherwise set} \; k_i^{(j)} = j \; \text{and} \; \delta_i^{(j)} = \beta_i^{(j)}; \]

[PF2] simulate \( V_i^{(j)} \) from \( q_{i+1} \cdot V_i^{(j)} \).

[PF3] assign particle \( V_i^{(j)} \) a weight

\[
 w_i^{(j)} = w_i^{(k_i^{(j)})} q_i^{(j)} f(y_{i+1} | V_i^{(j)}) p_{i+1} (V_i^{(j)} | V_i^{(k_i^{(j)})}) \beta_i^{(k_i^{(j)})} q_{i+1} (V_i^{(j)} | V_i^{(k_i^{(j)})}, y_{i+1}) .
\] (4.5)

end for

reduces to a direct importance sampling with target \( \pi_{i+1} \) 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 [12] and central limit theorems for estimates of posterior moments [14, 10, 32], as \( N \to \infty \). Under conditions relating to exponential forgetting of initial conditions for the signal, PF errors stabilise as \( n \to \infty \) [13, 32]. Additionally, the filtering distributions are obtained at computational cost \( O(N) \), 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 [9].

Algorithm 4.1 applies generally to state space time series models. However, when the signal is a discretely-sampled diffusion process, the PF cannot be applied due to the intractability of the system transition density, which is necessary in the calculation of the weights. One way to bypass this problem is to simulate the particles \( V_i \) according to the diffusion dynamics; then the transition density cancels out from Eq. (4.5). This requires the exact simulation of diffusions, which is discussed in Section 4.4. Another possibility is to try to obtain unbiased estimators for the transition density \( p_s(u, v) \) 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 [33].

Section 4.6 shows how for each pair \( (u, v) \) and times \( (s, t) \), with \( s < t \), to simulate auxiliary variables \( \Psi \) 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 \( \mathbb{E}[r(\Psi)] = p_s(u, v) \). Then, the so-called random weight PF (RWPF) inserts a further step between PF2 and PF3: simulate \( \Psi_i^{(j)} \) from \( Q(\cdot ; t_i, t_{i+1}, V^{(k_i^{(j)})}, V^{(j)}) \) and compute \( r(\Psi_i^{(j)}; t_i, t_{i+1}, V^{(k_i^{(j)})}, V^{(j)}) \). This quantity replaces the intractable transition density in Eq. (4.5). The RWPF is introduced in [46] and [21].

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 \( V_i^{(j)} \), \( \Psi_i^{(j)} \) is independent of \( \Psi_j \) 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 \( u, v, r(\psi, t, t_{i+1}, u, v) \) is
Chapter 4.7. Random Weight Continuous-Discrete Particle Filtering

4.2. Random Weight Continuous-Discrete Particle Filtering

a probability density function as a function of \((\psi, v)\) with respect to the product measure \(Leb(dv) \times Q(d\psi, t_0, t_1, t_{i+1}, u, v)\), 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 \(f(y_i | V_i)\). By construction the marginal filtering distributions of \(V_i\) in this model are precisely \(\pi_i\). Consider an auxiliary PF applied to this model where we choose with probabilities \(\beta^{(j)}_i\) each of the existing particles \((V^{(j)}_i, \Psi^{(j)}_i)\), 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, \(\Psi_{i+1}\) 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 PF on this discrete-time model whose latent structure has been augmented with the auxiliary variables \(\Psi_i\). It is worth mentioning that the potential of using unbiased estimators of intractable densities while retaining the “exactness” of MC algorithms is being increasingly recognised. The idea already appears in a disguised form in the auxiliary PF of [43] and explicitly in the rejection control algorithm, see for example Section 2.6.1 of [33]. [7] elaborate on this idea already appears in a disguised form in the auxiliary PF of [43] and explicitly in the reduction approach of [25] in the context of option pricing.

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

In the auxiliary PF formulation given in Algorithm 4.1 the positivity of the estimators is not necessary, since the resampling probabilities are controlled by \(\beta_i\). 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 \(w_j\)'s lose their interpretation as un-normalized probabilities; this is further discussed in Section 4.7. On the other hand, the generic algorithm in Algorithm 4.2, proposed originally in [22], 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 \(r^{(i,j)} = 1, 2, \ldots, \) 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 \(\tau\) is a stopping time which depends on the sign of all weights. If \(\mathbb{E}[\tau] < \infty\), then \(\mathbb{E}[r^{(j)} \mid w^{(j)}] = \mathbb{E}[\tau]w^{(j)}\); this follows from Wald’s identity, see Theorem 2 of [22]. The intractable normalizing constant \(\mathbb{E}[\tau]\) 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.

Algorithm 4.2 Creating positive unbiased importance weights exploiting Wald’s identity.

\[
\begin{align*}
\text{Set } i &= 1, \text{ simulate } r^{(1,j)} \text{ and set } r^{(j)} = r^{(1,j)}, \text{ for all } j = 1, \ldots, N. \\
\text{If } \min_j \{r^{(j)}\} > 0 \text{ then STOP.} \\
\quad i &= i + 1, \text{ simulate } r^{(i,j)} \text{ and set } r^{(j)} = r^{(j)} + r^{(i,j)}, \text{ for all } j = 1, \ldots, N. 
\end{align*}
\]
4.3 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 Eq. (4.1) is time-homogeneous.

(A1) In the SDE (4.1), \( d = m, \sigma = I, \) and \( b \) is of gradient form, i.e., there exists a function \( U : \mathbb{R}^d \rightarrow \mathbb{R} \) (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 \( v \rightarrow x = \int^v 1/\sigma(u)du \). 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:

\[
dX_s = \nabla U(X_s) \, ds + dB_s, \quad s \in [0,T];
\]

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 Eq. (4.6). We can appeal to the Cameron-Martin-Girsanov theorem for It\'o processes, see for example Theorem 8.6.6 of [38], 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

\[
\frac{d\mathbb{P}_b}{d\mathbb{P}_0}\big|_t(X) = \exp \left\{ U(X_t) - U(X) - \int_0^t \phi(X_s)ds \right\},
\]

where \( \phi(u) := (||b(u)||^2 + \Delta U(u))/2, \Delta \) 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 \( y \in \mathbb{R}^d \). 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 \( X_t \):

\[
\frac{d\mathbb{P}_b}{d\mathbb{P}_0}\big|_y(X) = \frac{p_{0,t}(x,y)}{\mathcal{G}_{0,t}(x,y)} \frac{d\mathbb{P}^{*}_{b_{t,y}}}{d\mathbb{P}^{*}_{0_{t,y}}}(X),
\]

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

\[
\frac{d\mathbb{P}^{*}_{b_{t,y}}}{d\mathbb{P}^{*}_{0_{t,y}}}(X) = \frac{\mathcal{G}_{0,t}(x,y)}{p_{0,t}(x,y)} \exp \left\{ U(y) - U(x) - \int_0^t \phi(X_s)ds \right\}.
\]

Re-arranging Eq. (4.8) and taking expectations on both sides with respect to \( \mathbb{P}^{*}_{0_{t,y}} \), we obtain the following representation for the transition density:

\[
p_{0,t}(x,y) = \mathcal{G}_{0,t}(x,y) \exp[U(y) - U(x)]\mathbb{E}_{\mathbb{P}^{*}_{0_{t,y}}} \left[ \exp \left\{ - \int_0^t \phi(X_s)ds \right\} \right].
\]
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 (4.8) can be made formal, see for example Theorem 2 of [15]. On a more general level, Eq. (4.9) follows from the basic principles of conditional expectation. In particular let \((\Omega, \mathcal{F})\) be a measurable space, \(P\) and \(Q\) be two probability measures on the space with Radon-Nikodym derivative \(\xi = dP/dQ\), and let \(\mathcal{G} \subseteq \mathcal{F}\) be a sub-\(\sigma\)-algebra. Then, the derivative \(dP/dQ\) restricted to \(\mathcal{G}\) is \(E[\xi | \mathcal{G}]\). 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 [54]), 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 in [16] it is used to define an EM algorithm for partially observed continuous-time Markov processes. To obtain Eq. (4.9) we specify \(\mathcal{G}\) as the \(\sigma\)-algebra generated by \(X_t\) and use the result in conjunction with Eq. (4.7).

For a thorough presentation of MC 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 [40].

### 4.4 Exact Simulation of Diffusions

[5] and [6] recognised that Eq. (4.7) 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 [5] and EA3 in [6].

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

(A2) The function \(\phi\) in Eq. (4.9) 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 \(\phi\) as

\[
\phi(u) = (||b(u)||^2 + \Delta U(u))/2 - \ell \geq 0.
\]

We fix a time horizon \(t\), such that (A3) holds, and consider the problem of simulating \(X_t\) according to the solution of Eq. (4.6) given \(X_0 = x\), or equivalently according to the transition distribution \(p_{0,t}(x,y)dy\). (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 \(P_0\),

\[
\frac{dZ}{dP_0}|_t(X) = \exp\{U(X_t)\},
\]

that is \(Z\) is obtained from \(P_0\) 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

\[
\frac{dP_b}{dZ}|_t(X) \propto \exp\left\{-\int_0^t \phi(X_s)ds\right\} \leq 1.
\]

(4.11)
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 \(P_h\) using proposals from \(Z\) and accepting them with probability (4.11). 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 [6].

**Theorem 4.1.** (Wiener-Poisson factorization) Let \(L\) denote the law of a unit rate Poisson process on \([0, t] \times [0, \infty)\) and define the extended law \(Z \otimes L\) with typical realisation \((X, \Phi)\), with \(\Phi = \{(\chi_j, \psi_j)\}_{j \geq 1}\), and \(|\psi_j|\) non-decreasing. Define the event,

\[
\Gamma := \bigcap_{j \geq 1} \{\phi(X_{\chi_j}) < \psi_j\}
\]

Then, \(P_h\) on \([0, t]\) is the marginal distribution of \(X\) when \((X, \Phi) \sim Z \otimes L|\Gamma\).

Effectively, the theorem formalizes the observation that the exponential term in Eq. (4.11) can be identified as the probability that an independent Poisson process on \([0, t] \times [0, \infty)\) has no points under the epigraph of \(s \rightarrow \phi(X_s)\) for a given path \(X\). The connection between the Poisson process and the diffusion measure is investigated and motivated further in Section 4.5.

Given knowledge of the range of \(\phi\) we can appeal to the principle of retrospective sampling [39] 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 \(\phi\) is also upper bounded,

\[
\text{there exists an } r < \infty \text{ such that } \sup_u \phi(u) < r.
\]

Then, the conditioned posed by Eq. (4.12) 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 \(\Phi\) is independent of \(X\), we can first simulate the Poisson process on \([0, t] \times [0, r]\) and unveil \(X\) at the times \(\chi_j\) specified by the Poisson process. When Eq. (4.12) 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 \((X, \Phi)\) according to \(Z \otimes L\) and accepting them when \((X, \Phi) \in \Gamma\), where we have used the upper bound of \(\phi\) and retrospective sampling to check the condition using finite computation. The algorithm is given in Algorithm 4.3, and each accepted draw \(X_t\) is a sample from the target diffusion at time \(t\). For convenience, the \(X_{\chi_j}\)'s are now ordered whereas the \(\psi_j\)'s are not. Note also that Step 2 simulates from the finite-dimensional distributions of \(Z\).

**Algorithm 4.3** The Exact Algorithm for the simulation of \(X_t\) according to the SDE (4.6) when Eq. (4.13) holds (EA1).

1. Generate a Poisson process \(0 < \chi_1 < \chi_2 < \cdots\) of rate \(r\) on \([0, t]\). Let \(\kappa\) be the number of points. Generate a sequence of uniform random variables \(\psi_j \sim \text{Uni}[0, r]\), \(j = 1, \ldots, \kappa\).
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 2.

When \(\phi\) is unbounded, the joint simulation according to \(Z\) of \(X\) and a random box which contains it, is required. This is the EA3 which is described in detail in [6]. The extra effort

\[
|\psi_j| < \infty
\]
needed in EA3 comes at an increased computational cost: the careful and extensive numerical investigation in [42] 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 \(O(t)\) 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 structure of the drift and the form of Eq. (4.11). This argument is supported empirically in [42], who find that EA1 has complexity \(O(d)\) 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.

### 4.5 Exact Simulation of Killed Brownian Motion

The Wiener-Poisson factorization in Theorem 4.1 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 (4.6) as killed Brownian motion; see for example Section 8.2 of [38] where also the connections to the Feynman-Kac formula are discussed. In particular, consider an exponentially distributed random variable \(E \sim \text{Exp}(1)\), independent of \(X\) and define the killing time \(T\) as the following function of \(E\) and \(X\):

\[
T = \inf \left\{ s : \int_0^s \phi(X_s) \, ds = E \right\},
\]

where \(\phi\) is given in Eq. (4.10). Thus,

\[
\Pr [T > t \mid X] = \exp \left\{ -\int_0^t \phi(X_s) \, ds \right\}.
\] (4.14)

Then, it is easy to see that the scheme described in Algorithm 4.4 yields an importance sampling approximation of the law of \(X_t\) induced by the SDE (4.6). The resulting weighted

**Algorithm 4.4** Importance sampling approximation of the law of \(X_t\) by killed Brownian motion.

1. Set \(j = 0\).
2. while \(j < N\) do
3. Generate \(E \sim \text{Exp}(1)\).
4. Generate a Brownian path \(X\) started from \(x\), and keep track of \(\int_0^s \phi(X_s) \, ds\). Stop when \(s = t\).
5. Rejection: If \(\int_0^s \phi(X_s) \, ds > E\) reject the path, goto 1.
6. Weighting: If \(\int_0^s \phi(X_s) \, ds < E\) then \(j := j + 1\), set \(X_t^{(j)} = X_t\), \(w_t^{(j)} = e^{U(X_t)}\). Goto 2.
7. end while

sample \(\{X_t^{(j)}, w_t^{(j)}\}_{j=1}^N\) is a particle approximation of the law of \(X_t\). The killing step (Step 4 in Algorithm 4.4) 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 Eq. (4.14), and the weighting (Step 5) is necessary to ensure that the path has density proportional to Eq. (4.7). However,
the scheme of Algorithm 4.4 is not practically implementable, since it involves an infinite amount of simulation in Step 3.

Note that for a given \( X, T \) is the first arrival time of a time-inhomogeneous Poisson process with intensity \( \phi(X) \). Assume now that Eq. (4.13) 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 \( \chi_j \) is accepted with probability \( \phi(X_{\chi_j})/r \), \( T \) is the first accepted arrival time. Algorithm 4.5 is a modification of Algorithm 4.4; we call it the Exact Killing (EK) algorithm. The

**Algorithm 4.5** Exact Killing: Exact simulation of a killed Brownian motion using thinning.

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

resulting 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 (4.13).

Algorithm 4.5 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 [23] 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 Algorithm 4.3. 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 \( t \) a priori, 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 \( t \to \mathbb{P}_{bl}. \) Since EK also relies on rejection sampling, the computational effort to yield a particle \( t \) increases exponentially with \( t \). The Markov property can be exploited here as well, by defining time increments of size, \( \delta \) say. If a particle is alive at time \( t \delta \) 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 \( O(t) \).

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 \( \chi_j \), 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 $\psi_i < \phi(X^{(i)}_t)$. 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 $P_{bi}$ we weight each alive particle $X^{(j)}_t$ at time $t$ with $w^{(j)}_t = e^{U(X^{(j)}_t)}$ weight. Hence, we can simulate exactly the genealogy of this interacting particle systems which tracks the law of the diffusion process.

4.6 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 MC solution to fixed point problems, see for example [28] for applications in population genetics, [52] in the context of solutions of Partial Differential Equations (PDEs), [18] for a recent contribution in the literature and references, and Section 2.5.6 of [33] 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 8.6. 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 4.6.2. 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 4.6.3, 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 4.6.4, and provides the optimal importance sampling estimator for the problem posed in Section 4.6.3. This formalism links directly with the so-called MC method for solving integral equation and fixed-point problems. This is outlined in Section 4.6.4. 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 4.6.6.

4.6.1 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 $E[\tilde{X} \mid X] = X$ (we will use $\tilde{X}$ to
Monte Carlo Probabilistic Inference for Diffusion Processes

denote a generic element of the sequence). We will also assume that the $\tilde{X}_j$s have a common finite absolute moment, $E[|\tilde{X}| \mid X] < \infty$. 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) $E[f(X)]$ 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{align}
e^X &= e^c \sum_{i=0}^{\infty} \frac{(X - c)^i}{i!} \\
 &= e^c \sum_{i=0}^{\infty} E \left[ \prod_{j=1}^{i} (\tilde{X}_i - c) \mid X \right] / i! = e^c \left[ \sum_{i=0}^{\infty} \prod_{j=1}^{i} (\tilde{X}_i - c) / i! \right] 
\end{align}

(4.15)

where the product $\prod_{j=1}^{0} j$ is defined to be equal to 1. The role of $c$ will be discussed later. Note that the absolute moment assumption on the $\tilde{X}_j$s 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 $\lambda > 0$,

\begin{align}
e^{1+c} \prod_{j=1}^{\kappa} \frac{\tilde{X}_j - c}{\lambda}, \quad \kappa \sim Po(\lambda)
\end{align}

(4.16)

is a realizable unbiased estimator of $e^X$. We term Eq. (4.16) the *Poisson estimator*. Its second moment is easy to work out provided that the $\tilde{X}_j$s have a common second moment, $E[\tilde{X}^2 \mid X] < \infty$:

\begin{align}
\exp \left\{ \lambda + 2c + \frac{1}{\lambda} E[(\tilde{X} - c)^2 \mid X] \right\}
\end{align}

(4.17)

The two constants $c$ and $\lambda$ 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 4.2). 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 4.1.** 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 Eq. (4.17) we have that for fixed $c$, the optimal choice for $\lambda$ is $E[(\tilde{X} - c)^2 \mid X]^{1/2}$, whereas for a given computational budget $\lambda$, $c$ 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 $\lambda$ 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 4.6.3 and 4.6.6.
In most cases of interest $X$ is a random variable and we are actually interested in estimating $E[e^X]$ with respect to the law of $X$. The argument presented earlier can be repeated to show that Eq. (4.16) is unbiased for this quantity, however we need the stronger condition

$$E[\exp[\mathbb{E}[|\tilde{X} | X]]] < \infty$$

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

$$E[\exp\left\{\frac{1}{\lambda}(E[X^2 | X] - 2cX)\right\}] < \infty;$$

then, the expected value of Eq. (4.17) 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 $\lambda M$ and the choice of optimal allocation in terms of $\lambda$ and $M$ is non-trivial.

Furthermore, $c$ and $\lambda$ can be chosen to depend on $X$. [21] 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.

### 4.6.2 Unbiased Truncation of Infinite Series

In the previous section an estimator was given in terms of an infinite sum in Eq. (4.15). 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

$$S = \sum_{k=1}^{\infty} \alpha_k,$$  \hfill (4.18)

where we assume that the sum is finite a.s. As in the previous section we might be interested in $E[S]$ when the $\alpha_k$'s are random variables, but the argument follows in a similar way. There are (at least) three ways to obtain an unbiased estimator of Eq. (4.18), 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 $\alpha_K / \beta_K$ is an unbiased estimator of $S$, where $K$ is simulated according to $\Pr[K = k] = \beta_k$. If

$$S_a = \sum_{k=1}^{\infty} |\alpha_k| < \infty,$$  \hfill (4.19)

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 4.6.3), 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,

$$\frac{\alpha_K}{\prod_{i=1}^{K-1} (1 - p_i)p_K}$$
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 $\beta_k$, 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 Eq. (4.19). Let again $\beta_k$ be probabilities with survival function $G$. Then,

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

which suggests the following unbiased estimator of $S$:

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

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 (4.16) 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 Eq. (4.15) and apply a technique for the unbiased estimation of the infinite sum.

### 4.6.3 Unbiased Estimation of the Expected Value of Exponential Functions of Markov Process Path Integrals

A very interesting instance of the generic context of Section 8.6 is when $X$ is a path integral of a Markov process. With a slight abuse of notation, suppose that we are interested in estimating

$$
I(x, t) := E\left[\exp\left\{\int_{t}^{1} g(s, X_s) ds\right\}\right], \quad t \leq 1
$$

where $X$ is a Markov process in $\mathbb{R}^d$, with explicit transition density $p_{x,t}(x, y)$, 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 [51] 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 Eq. (4.20) is raised when considering the estimation of the transition density for the class of diffusion processes considered in Section 4.3; see Eq. (4.9) where $X$ is the Brownian bridge. The use of the estimators for the estimation of diffusion transition densities was considered in [7], see also Section 4.6.6.

By the standard MC integration trick, we have that $(1-t)g(\chi, X_t)$ where $\chi \sim \text{Uni}(t, 1)$, is conditionally on $X$ an unbiased estimator of the exponent in Eq. (4.20). Working precisely
as in Section 8.6, under the sufficient condition

\[ I_\alpha(x,t) := \mathbb{E} \left[ \exp \left\{ \int_t^1 |g(s,X_s)| ds \right\} \right] < \infty, \quad \text{for all } t \leq 1 \]

we get the following infinite-series unbiased estimator of Eq. (4.20):

\[
\sum_{k=0}^{\infty} \int_t^1 \cdots \int_t^1 \int_t^1 \cdots \int_t^1 \prod_{i=1}^{n} g(u_i,x_i)p_{u_i-1} d(x_{i-1}, x_i) dx_n \cdots dx_1 du_n \cdots du_1, \quad (4.21)
\]

with the convention that \( x_0 = x, u_0 = t \). This infinite expansion can be treated with the machinery of Section 4.6.2 to yield feasible unbiased estimators of Eq. (4.20). 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

\[
e^{(\lambda+c)(1-t)} \prod_{j=1}^{k} \frac{g(X_j, X_{j-1}) - c}{\lambda}, \quad \kappa \sim Po(\lambda(1-t)), X_j \sim \text{Uni}(t,1). \quad (4.22)
\]

Note however that with the same variables we can consider the alternative estimator based on the application of Fubini’s theorem discussed in Section 4.6.2, or indeed use a different proposal distribution for the index \( K \) (e.g. the negative binomial).

The specific structure of the exponent in Eq. (4.20) (as opposed to the generic one in Section 8.6) permits a mathematically richer formulation of the estimation problem. This is done in [51] (see in particular Propositions 1,2 and 4 of the article). This formulation casts the estimation of Eq. (4.20) as a familiar problem in MC. 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 \( X \) is the space \([t, 1] \times \mathbb{R}^\ell\). Let us now define the following signed measure \( \varphi \) on \( \mathcal{Y} \) indexed by \((x,t)\), and given by the formulas

\[
\varphi(k, d(t_0,x_0) \times \cdots \times d(t_k, x_k); x, t) = \delta_{(t,x)}(d(t_0,x_0))
\]

\[
\prod_{i=1}^{k} 1_{[t_{i-1},1]} g(t_i,x_i)p_{t_{i-1}} d(t_i, x_i)
\]

where \( \delta \) denotes the Dirac delta function, and \( 1_{[A]} \) is 1 if \( x \in A \) and 0 otherwise. In this formulation, Eq. (4.21) shows that Eq. (4.20) is the normalising constant of \( \varphi \): \( I(x,t) = \varphi(\mathcal{Y}; x, t) \), hence can reformulate the original 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 Eq. (4.20) as the computation of a normalising constant of a signed measure. The material of Section 4.6.2 together with standard MC techniques effectively it gives methods for constructing proposal distributions on \( \mathcal{Y} \) to be used in the importance sampling. [51] gives the following generic estimator where let \( p_0(s,x) > 0 \) is a killing probability and \( q_{x,y}(x,y) \) is an alternative tractable transition density:

\[
\prod_{i=1}^{K} \frac{g(x_{i+1}, x_{i})p_{x_{i+1}}(x_{i+1}, x_{i})}{(1 - p_0(x_{i+1}, x_{i}))q_{x_{i+1}}(x_{i+1}, x_{i})},
\]
where the $\chi_i$'s are ordered uniforms on $[t, 1]$ and the $X_{\chi_i}$ are generated according to the transitions $q$. Let $|\varphi|$ be the total variation of $\varphi$, thus it is obtained by replacing $g$ with its absolute value in the definition given above. Then, by Jensen’s inequality (as in Section 4.6.2) it follows that the optimal proposal distribution in terms of minimising the variance of the estimator, is $|\varphi|/I_d(x, t)$. 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 8.6) are available. For example

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

with the random elements as in Eq. (4.22); alternative numerical approximation of the integral in the exponents can be considered. A comparison among different schemes is carried out in Section 4.6.6.

### 4.6.4 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 4.6.3, where $\mathcal{X} = [t, 1] \times \mathbb{R}^d$.

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

$$\nu(k, dx_1 \times \cdots \times dx_{k+1} ; x) := \delta_{x_0}(dx_0) \prod_{i=1}^{k} p(x_{i-1}, x_i) \gamma(x_i) dx_1 \cdots dx_{k+1}.$$  

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 4.6.3. 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:

$$I(x) = 1 + \int_{\mathcal{Y}} I(x_1)p(x, x_1)\gamma(x_1)dx_1. \quad (4.23)$$

Using the same argument that lead to Eq. (4.23) we can obtain the following marginal-conditional distributions under $\tilde{\nu}$: $\tilde{\nu}(k = 0; x) = 1/I(x)$, $\tilde{\nu}(dx_1, k > 0; x) \propto p(x, x_1)\gamma(x_1)I(x_1)dx_1$. In the same way we obtain the general expressions

$$\tilde{\nu}(k > i - 1, dx_1, \ldots, dx_i ; x) = I(x_i) \prod_{j=1}^{i} p(x_{j-1}, x_j)\gamma(x_j)$$

$$\tilde{\nu}(dx_1 | x, x_1, \ldots, x_i, k > i - 1) = p(x_{i-1}, x_i)\gamma(x_i)I(x_i)dx_1/(I(x_{i-1}) - 1)$$

$$\tilde{\nu}(k = i | x, x_1, \ldots, x_i, k > i) = 1/I(x_i).$$
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 [18] 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.

### 4.6.5 Monte Carlo for Integral Equations

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

\[
I(x) = h(x) + \int_X p(x, y)I(y)dy
\]

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 Eq. (4.20) satisfies such an equation with \( h = 1 \). By successive substitution of \( I \) in the equation we obtain the infinite expansion

\[
I(x) = h(x) + \sum_{k=1}^{\infty} \int_{X^0} \prod_{i=1}^{k} p(x_{i-1}, x_i)h(x_k)dx_1 \times \cdots \times dx_k
\]

with the convention \( x_0 = x \). The analogy with the problems treated in Sections 4.6.2 and 4.6.4 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 8.6 do not necessarily lead to a fixed point problem. However, the techniques of Section 4.6.2 still apply to yield unbiased estimators even in these cases.

### 4.6.6 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 [11]. This is a one-dimensional diffusion with \( b \) and \( \sigma \) in Eq. (4.1) given by \( -\rho(x - \mu) \), and \( \sigma \sqrt{x} \) respectively, where \( \rho > 0, \sigma > 0, \mu \) are parameters and \( x \in \mathbb{R}^+ \). This diffusion is not in the form (4.6) but it can be transformed as described in Section 4.3. 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 [7]. However, in this model when the process is transformed its measure is absolutely continuous with respect to the law of the Brownian motion conditioned to remain positive, which is known as the Bessel process. Therefore, Eq. (4.9) 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 (4.22) to estimate the expectation in Eq. (4.9) (using Bessel bridge dominating measure). Second, a biased estimator based on Riemmann approximation of the exponent in Eq. (4.9). This is
in the spirit of the plug-in estimators discussed in Section 4.6.3 but where the times to evaluate the path are chosen deterministically. This estimator in the context of diffusions was considered in [37]. Finally, we consider two estimators obtained using the discrete-time approach of [19]. We use their estimator on the original CIR process and on the transformed to unit-diffusion-coefficient process. The estimator of [19] applied to the transformed process is closely related to the estimator of [37]: 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 [19]: \((\rho,\mu,\sigma) = (0.5,0.06,0.15)\) 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 estimators 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 [47]. For the estimation of the root mean square error of each estimator we average 120 independent replicates.

Fig. 4.1 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 Eq. (4.9). 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.

4.7 Discussion and Directions

We have reviewed and developed a rich methodological framework for the MC 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 [7]. It is interesting to understand deeper the connections; this might lead to new exact simulation algorithms outside the framework described in Section 4.4. A different instance of this interplay appears in the exact simulation from the stationary distribution of a Markov chain. There, a uniform ergodicity condition leads to an infinite-series expansion for the stationary distribution which can then be used for exact simulation.

The methodology for diffusions is based on the convenient representation of the transition density in Section 4.3, which relies on certain assumptions about the drift and the diffusion coefficient of the process. The conditions are strong when \(d > 1\). On the other hand, Theorem 3.1 of [52] establishes that the transition density of a generic diffusion process solves an integral equation of the type (4.24) with \(h\) and \(p\) explicitly given. This representation relies on different conditions which relate to smoothness and boundedness
4.7. Discussion and Directions

Figure 4.1: 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 $t = 1/252$ (left) and $t = 1/2$ (right). $\circ$ Durham and Gallant without variance transformation, $\triangle$ Durham and Gallant with variance transformation, (+) Nicolau, (x) Poisson Estimator.

of the drift and diffusion coefficients. Additionally, $p$ might not be a positive kernel. Nevertheless, this alternative representation is amenable to the type of unbiased MC estimation using the tools of Section 4.6, and this has been pursued in [52] and [50]. In current work we are exploring the possibilities of using this representation to achieve exact simulation of the diffusion at fixed times, and evaluating the practical usefulness of this alternative.

The article has given little attention to the important question of choosing between unbiased and biased estimators of the transition density and other diffusion functionals. This question has not been seriously addressed since the primary purpose of the article is to present in a unified manner a collection of ideas central to the construction of unbiased MC schemes. Biased estimators of diffusion functionals can be easily obtained using the Euler or other type of approximations; see the discussion in Sections 4.6.3 and 4.6.6. It is difficult to give very general statements about which type of estimator should be preferred, particularly since exact calculation of mean square error is complicated even in simple examples. Research in obtaining some simple general rules is underway. In the article we provide a simple comparison in Section 4.6.6 for the estimation of the CIR transition density, and a much broader evaluation of competing biased and unbiased MC schemes is in progress. On the other hand, a certain amount of empirical comparisons has been published, see for example Sections 4.1 and 5.1 of [21] and Section 4 of [50]. Wagner finds significant reduction in mean square error via the application of variance reduction
techniques and recommends a combination of unbiased estimators with such techniques. Closing this discussion, a generic argument in favour of unbiased estimation of unknown quantities within a MC scheme is that of “consistency”. The RWPF of Section 4.2 provides consistent estimates of the filtering distributions as $N \to \infty$. Working with biased estimates requires that the bias is eliminated at a specific rate hence consistency is achieved by letting both $N$ and the amount of imputation going to infinity at appropriate rates. Similarly, a MCEM algorithm will typically give consistent parameter estimates as the number of data go to infinity even with fixed MC effort. This is not so when the MC contains bias. The (appropriate) replacement of intractable densities by positive unbiased estimators within an MCMC algorithm [4, 3] does not perturb the limiting distribution; this is crucial since it is typically difficult to quantify the amount of bias that would be introduced otherwise.

Finally, as pointed out earlier, variance reduction techniques can be very effective in estimation of diffusion functionals. The biased Wiener measure proposal of Section 4.4 can be used for this purpose; see for example [50] for implementation of such ideas. Variance reduction methods for diffusions is studied for example in [36]. These possibilities within the exact simulation framework are being currently investigated.

Acknowledgments The author would like to acknowledge financial support by the Spanish government through a “Ramon y Cajal” fellowship and the grant TM2008-06660, 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.

Bibliography


111


