Bernoulli **23**(3), 2017, 2028–2057
DOI: 10.3150/16-BEJ803

Unbiased simulation of stochastic differential equations using parametrix expansions

PATRIK ANDERSSON[1] and ARTURO KOHATSU-HIGA[2]

[1] *Department of Statistics, Uppsala University, Box 513, 751 20 Uppsala, Sweden. E-mail: patrik@ndersson.nu*

[2] *Department of Mathematical Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan. E-mail: arturokohatsu@gmail.com*

In this article, we consider an unbiased simulation method for multidimensional diffusions based on the parametrix method for solving partial differential equations with Hölder continuous coefficients. This Monte Carlo method which is based on an Euler scheme with random time steps, can be considered as an infinite dimensional extension of the Multilevel Monte Carlo method for solutions of stochastic differential equations with Hölder continuous coefficients. In particular, we study the properties of the variance of the proposed method. In most cases, the method has infinite variance and therefore we propose an importance sampling method to resolve this issue.

Keywords: importance Sampling; Monte Carlo method; multidimensional diffusion; parametrix

1. Introduction

Consider the following multidimensional stochastic differential equation (s.d.e.)

$$X_t = X_0 + \sum_{j=1}^{m} \int_0^t \sigma_j(X_s)\, dW_s^j + \int_0^t b(X_s)\, ds, \qquad t \in [0,T]. \tag{1}$$

Here, W is an m-dimensional Wiener process and $\sigma_j, b : \mathbb{R}^d \to \mathbb{R}^d$ are such that there exists a weak solution to (1). More precise assumptions that we will work under and that guarantees this will be stated later.

As there are very few situations where one can solve (1) explicitly one has to use numerical approximations. In such a case, the numerical calculation of $\mathsf{E}[f(X_T)]$ for $f : \mathbb{R}^d \to \mathbb{R}^d$, using Monte Carlo simulation has been widely addressed in the literature. In particular, the Euler–Maruyama scheme is one of the main numerical approximation schemes studied due to its generality and simplicity of implementation.

Given a time partition $\pi : 0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = T$, we define the Euler–Maruyama scheme $\{X_{t_i}^{\pi}; i = 0, \ldots, N+1\}$ associated with (1) by $X_0^{\pi} = X_0$,

$$X_{t_{i+1}}^{\pi} = X_{t_i}^{\pi} + \sum_{j=1}^{m} \sigma_j\bigl(X_{t_i}^{\pi}\bigr)\bigl(W_{t_{i+1}}^j - W_{t_i}^j\bigr) + b(X_{t_i})(t_{i+1} - t_i), \qquad i = 0, \ldots, N.$$

1350-7265 © 2017 ISI/BS

From this numerical approximation, it is possible to draw Monte Carlo samples and therefore we can obtain an approximation of $\mathsf{E}[f(X_T)]$. This approximation will contain two types of errors, a statistical error which arises from the fact that we are taking the average of a finite number of Monte Carlo samples. This error can usually be controlled by estimating the variance of the sample. Should a smaller error be needed we simply draw more Monte Carlo samples. The second error is the bias, which comes from the time discretization π. This error is more difficult to control. There are results on the asymptotic rate at which this error decreases as the number of steps increases but in general it is not possible to know (a priori) how large this error is in a specific example. In cases where the coefficients have some regularity the rate of convergence is known (see [3]).

Although these convergence rates may be considered to be slow, they have become the basis of the construction of other more refined numerical schemes. Besides, this computational application, they also have various theoretical uses and is therefore important.

Many researchers have addressed various of the above issues related to some of the shortcomings of the Euler–Maruyama scheme. In order to carry out these studies, one needs in general smoothness and uniform ellipticity assumptions on the coefficients of (1).

In this article, we provide and discuss the properties of a numerical scheme with no bias and which is also applicable to situations where the coefficients σ or b are bounded Hölder continuous functions and therefore they may not satisfy smoothness conditions.

For (weak) existence and uniqueness of solutions of s.d.e.'s with bounded Hölder continuous uniformly elliptic coefficients, we refer the reader to [18].

The organization of this paper is as follows. After introducing some definitions and notations in the next section, we follow in Section 3, introducing the Multilevel Monte Carlo (MLMC) methods and in particular the parametrix method which can be seen as a random MLMC method. In Section 4, we derive some bounds that will be useful in understanding the behavior of the variance of the proposed simulation method. Then in Sections 5 and 6, we explore some of the properties of the two varieties of the parametrix method, in particular we find that the simulation methods do not always give a finite variance.

In Section 7, we show how, using importance sampling on the discretization times of the Euler–Maruyama scheme, the problem of infinite variance can be solved. The main results are that, assuming that the coefficient functions are regular, the method achieves finite variance, and that, assuming only that the coefficient functions are Hölder, the method achieves finite variance in dimension 1. In Section 8, we find bounds on the variance of the methods and we are therefore able to state an optimization problem for finding parameters in the importance sampling method which will improve the performance of the simulation methods. We also provide some rules of thumb for choosing these parameters. In Section 9, we exemplify the results obtained in the paper by applying these methods to some s.d.e.'s. We find that our examples behaves as we would expect from the results developed in this article.

2. Definitions and notations

Here we give some of the notations and definitions that will be used throughout the text. For two symmetric matrices a and b, we let $a < b$ mean that $b - a$ is positive definite. Also let

$a^{i,j}$ denote the (i,j)th element of a. Let a be a $d \times d$ symmetric non negative definite matrix, with $0 < \underline{a}I \leq a \leq \overline{a}I$, for $\underline{a}, \overline{a} \in R$, where I is the $d \times d$ dimensional identity matrix. Define $\rho_a := \overline{a}/\underline{a}$. The multi-dimensional Gaussian density with mean zero and covariance matrix a is denoted by

$$\varphi_a(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det a}} \exp\left\{-\frac{1}{2}x^T a^{-1} x\right\}.$$

We abuse the notation by using $\varphi_{\overline{a}} \equiv \varphi_{\overline{a}I_{d\times d}}$. We let ∂_α^j be the partial derivative operator of order j, with respect to the variables

$$\alpha = (\alpha_1, \dots, \alpha_j) \in \mathcal{A}_j = \left\{\alpha = (\alpha_1, \dots, \alpha_j) \in \{1, \dots, d\}^j\right\}$$

and define the Hermite polynomials associated with the Gaussian density, $H_a^i(x) = -(a^{-1}x)^i$ and $H_a^{i,j}(x) = (a^{-1}x)^i(a^{-1}x)^j - (a^{-1})^{i,j}$. That is, $\partial_i \varphi_a(x) = H_a^i(x)\varphi_a(x)$ and $\partial_{i,j}^2 \varphi_a(x) = H_a^{i,j}(x)\varphi_a(x)$. In general, $|\cdot|$ denotes the norm in real vector spaces while $\|\cdot\|_k$ denotes the uniform norm in $C_b^k(\mathbb{R}^d, \mathbb{R}) \equiv C_b^k(\mathbb{R}^d)$, the space of bounded functions with k bounded derivatives. The norm in this space is defined

$$\|f\|_k = \sum_{j=0}^{k} \sum_{\alpha \in \mathcal{A}_j} \sup_{x \in \mathbb{R}^d} \left|\partial_\alpha^j f(x)\right|.$$

$C_c^\infty(\mathbb{R}^d)$ denotes the space of real valued infinitely differentiable functions with compact support defined on $\mathbb{R}^d$. For a bounded function f, recall that $\|f\|_0 = \sup_{x \in \mathbb{R}^d} |f(x)|$. For a Hölder function $a : \mathbb{R}^d \to \mathbb{R}^k$ with index $\alpha \in (0,1]$, we define its Hölder norm as $a_H = \sup_{x \neq y} \frac{|a(x)-a(y)|}{|x-y|^\alpha}$. As above, we naturally extend the definition so that $C_b^\alpha(\mathbb{R}^d, \mathbb{R}) \equiv C_b^\alpha(\mathbb{R}^d)$ denotes the space of bounded α-Hölder functions. $\delta_{x_0}(x)$ will denote the Dirac's delta generalized distribution function.

Throughout the proofs, we will use a constant denoted by C_T in order to indicate the dependence on T. This constants may change value from one line to the next. Furthermore, they will always be increasing in T and converge to a finite value as $T \downarrow 0$.

3. Multilevel Monte Carlo methods

Because of the difficulty to quantify the discretization error, there is an interest in so-called unbiased simulation methods. We differentiate here between exact methods and unbiased methods. Exact meaning to sample a path, at a finite set of points, with the distribution of the s.d.e., while unbiased means that we can, with out bias, estimate $\mathsf{E}[f(X_T)]$. In many applications, an unbiased method is however sufficient. An example of an exact method is given in [4]. This method uses the Lamperti transform of the original s.d.e. and therefore it is limited to the case $d = m = 1$. Other exact methods have also been derived for some special cases, for example, [6] for the SABR model and [5] for the Heston model. However, these methods can not be easily extended to a general s.d.e.

Now, in order to introduce the Multilevel Monte Carlo method (MLMC), let X_T^n denote an approximation of X_T using an Euler method with uniform time steps of length $2^{-n}T$, $n \geq 0$. The MLMC method, introduced in [10], is then to approximate $E[f(X_T)]$,

$$\mathsf{E}[f(X_T)] \approx \mathsf{E}[f(X_T^0)] + \sum_{n=1}^{\bar{n}} \mathsf{E}[f(X_T^n) - f(X_T^{n-1})],$$

up to some final level $\bar{n}$. At each level, a certain number of Monte Carlo samples are generated and the sample average then approximates the expectation. Choosing the number of samples at each level in a good way will improve the convergence rate compared to the Euler method.

Now, letting $\bar{n} \to \infty$ we can write, assuming convergence,

$$\begin{aligned}\mathsf{E}[f(X_T)] &= \mathsf{E}[f(X_T^0)] + \sum_{n=1}^{\infty} \mathsf{E}[f(X_T^n) - f(X_T^{n-1})] \\ &= \mathsf{E}[f(X_T^0)] + \sum_{n=1}^{\infty} p_n \frac{\mathsf{E}[f(X_T^n) - f(X_T^{n-1})]}{p_n} \\ &= \frac{1}{p_0}\mathsf{E}[1(N=0)f(X_T^0)] + \mathsf{E}\left[1(N \geq 1)\frac{f(X_T^N) - f(X_T^{N-1})}{p_N}\right],\end{aligned} \tag{2}$$

where N is a random variable with distribution $p_n > 0$, $n \geq 0$. The last expression above can be simulated unbiasedly using Monte Carlo methods. Now, we would like to discuss the second moment of the above proposed estimator. A straightforward calculation gives

$$\mathsf{E}\left[\left(\frac{f(X_T^N) - f(X_T^{N-1})}{p_N}\right)^2\right] \leq C \sum_{n=1}^{\infty} \frac{r_n}{p_n},$$

where $r_n := \mathsf{E}[(f(X_T^n) - f(X_T^{n-1}))^2]$. As discussed previously, the rate at which this error goes to zero is well understood. In fact, under sufficient regularity hypotheses on b, σ and f one knows that $r_n = O(2^{-n})$. Then the variance of the method will be finite if we choose $p_n \sim 2^{-n}n^2$. On the other hand, note that the average number of random number generators used can be considered to be $\sum_{n=1}^{\infty} 2^n p_n = \infty$. Therefore the procedure is doomed to fail as long as ∞-level Monte Carlo method goes. The choice of p_n may be changed in order to make the average complexity finite but then the variance of the method will be infinite.

The interpretation of the method is clear. The method has no bias because it relies on an infinite order expansion. The level used in each simulation is determined by the value of the random variable N and the amount of simulation in each level is determined by the choice of the probability distribution $p_n, n \geq 0$. Still, as it can not be applied in the ∞-dimensional case, one tries to approximate it by taking a certain number of levels. In fact, the previous calculation shows that taking too many levels in the MLMC may lead to the wrong result. One remedy is to use the Milstein scheme which improves on the order of strong convergence. However, this assumes more regularity and is also difficult to use in multidimensional problems (for more on this, see [11]).

Furthermore, in the case that the coefficients σ, b are Hölder functions and f is not a regular function it is well known that the rate r_n degenerates quickly (see, e.g., [14] and [1]). Therefore, the applicability of the MLMC method as understood above is limited. For this reason, we will propose to use the parametrix method of approximation as one possible extension of the MLMC method in what follows. This method will allow a more profound analysis of the differences between approximations showing exactly where the variance explosion appears. This variance problem will then be solved by an appropriate time importance sampling. Then an optimization procedure for the efficiency of the algorithm can be carried out.

The principle of simulation without bias just described can also be found in [16] and [17] which have already appeared in [15] which cites [12] as a source of this idea. As explained in Section 3 of [17], one can not apply $L^2(\Omega)$ criteria to this problem and even if a criteria in probability is applied as in Section 4 in that same paper then the computational complexity increases as the strong rate of convergence slows down.

3.1. Parametrix methods

The unbiased simulation technique for the multidimensional s.d.e. (1) which we will propose here has been introduced in [2]. This technique is based on the parametrix method introduced by E. Levi more than 100 years ago in order to solve uniformly elliptic partial differential equations of parabolic type. This method is highly flexible and has been extended to various other equations. The proposed method is also based on the Euler scheme although in this new simulation scheme the partition will be random.

We may interpret the method as a randomized MLMC method but where the structure of the s.d.e. is used to rewrite the difference of levels in (2) so that we can eventually handle Hölder type coefficients. In fact, heuristically speaking, the difference between two levels in (2), $f(X_T^N) - f(X_T^{N-1})$, can be rewritten using the Itô formula. This will generate a weight which appears due to the derivatives in the Itô formula as well as the difference of the generators of the two processes X^N and X^{N-1}. The differential operators which appear in the difference of the two generators can be applied to densities directly or in an adjoint form which therefore does not require differentiation of the coefficients. This requires a delicate "diagonal" type argument which appears in [2]. Finally, these arguments lead to two different types of approximations (for more details, see [2]). For this reason, one is called the forward method which requires smoothness of coefficients and the backward method which can be applied for Hölder continuous coefficients.

Let us introduce these methods in the following general format:

Let $\pi : 0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = T$ and define the following discrete time process and its associated weight function,

$$\begin{aligned}
&X_0^\pi \text{ is random variable with density } \nu(x),\\
&X_{t_{i+1}}^\pi = X_{t_i}^\pi + \mu\big(X_{t_i}^\pi\big)(t_{i+1} - t_i) + \sigma\big(X_{t_i}^\pi\big)\sqrt{t_{i+1} - t_i} Z_{i+1}, \qquad i = 0, 1, \dots, N,\\
&Z_i, i = 1, \dots, N \text{ are independent } \mathsf{N}(0, I_{m\times m}) \text{ random vectors},\\
&\theta_t(x, y) = \frac{1}{2}\sum_{i,j=1}^{d} \kappa_t^{i,j}(x, y) - \sum_{i=1}^{d} \rho_t^i(x, y).
\end{aligned} \tag{3}$$

We shall abuse this notation slightly and write for example, $X^{\pi}_{s_i}$ or $X^{\pi}_{\tau_i}$ with the understanding that the time partition π is appropriately defined. That is $\pi : 0 = s_0 < s_1 < \cdots < s_N < s_{N+1} = T$ or $\pi : 0 = \tau_0 < \tau_1 < \cdots < \tau_N < \tau_{N+1} = T$, which should be understood from the context.

Here, we define $a(x) \equiv \sigma^T \sigma(x)$ and assume that a is uniformly elliptic. As explained previously, the goal is to give an alternative probabilistic representation for $\mathsf{E}[f(X_T)]$ for $f : \mathbb{R}^d \to \mathbb{R}^d$.

Define $S^n = \{s = (s_1, \dots, s_n) \in R^n | 0 < s_1 < s_2 < \cdots < s_n < T\}$. Then, the following is proved in [2],

$$\mathsf{E}\big[f(X_T)\big] = \sum_{n=0}^{\infty} \int_{S^n} \mathsf{E}\left[\Phi\big(X^{\pi}_T\big) \prod_{j=0}^{n-1} \theta_{s_{j+1}-s_j}\big(X^{\pi}_{s_j}, X^{\pi}_{s_{j+1}}\big)\right] ds, \tag{4}$$

where we define $\int_{S^0} ds \equiv 1$. Now, let $N(t)$ be a Poisson process with intensity parameter $\lambda > 0$ and define $N \equiv N(T)$. Let $\tau_1, \dots, \tau_N$ be the event times of the Poisson process and set $\tau_0 = 0$, $\tau_{N+1} = T$. Since, conditional on N, the event times are distributed as a uniform order statistic, $P(N = n, \tau_1 \in ds_1, \dots, \tau_n \in ds_n) = \lambda^n e^{-\lambda T}$, for $s \in S^n$. We may rewrite the time integral in (4) in a probabilistic way as

$$\mathsf{E}\big[f(X_T)\big] = e^{\lambda T} \mathsf{E}\left[\Phi\big(X^{\pi}_T\big) \prod_{i=0}^{N-1} \lambda^{-1} \theta_{\tau_{i+1}-\tau_i}\big(X^{\pi}_{\tau_i}, X^{\pi}_{\tau_{i+1}}\big)\right]. \tag{5}$$

Now, the forward method is defined by

$$\left.\begin{aligned}
&\nu(x) = \delta_{X_0}(x),\\
&\Phi(x) = f(x),\\
&\mu(x) = b(x),\\
&\kappa^{i,j}_t(x, y) = \partial^2_{i,j} a^{i,j}(y) + \partial_j a^{i,j}(y) H^i_{ta(x)}\big(y - x - b(x)t\big)\\
&\qquad\qquad + \partial_i a^{i,j}(y) H^j_{ta(x)}\big(y - x - b(x)t\big)\\
&\qquad\qquad + \big(a^{i,j}(y) - a^{i,j}(x)\big) H^{i,j}_{ta(x)}\big(y - x - b(x)t\big),\\
&\rho^i_t(x, y) = \partial_i b^i(y) + \big(b^i(y) - b^i(x)\big) H^i_{ta(x)}\big(y - x - b(x)t\big).
\end{aligned}\right\} \tag{F}$$

Here, H^i and $H^{i,j}$ denote the Hermite polynomials of order 1 and 2 which have been defined in Section 2. Further assuming that $f(x)$ is a density function, the backward method is

$$\left.\begin{aligned}
&\nu(x) = f(x),\\
&\Phi(x) = \delta_{X_0}(x),\\
&\mu(x) = -b(x),\\
&\kappa^{i,j}_t(x, y) = \big(a^{i,j}(y) - a^{i,j}(x)\big) H^{i,j}_{ta(x)}\big(y - x + b(x)t\big),\\
&\rho^i_t(x, y) = \big(b^i(x) - b^i(y)\big) H^i_{ta(x)}\big(y - x + b(x)t\big).
\end{aligned}\right\} \tag{B}$$

We note here the time directional nature of each scheme from which the names forward and backward come from. In particular, in the backward method one needs to evaluate the irregular function $\Phi(x) = \delta_{X_0}(x)$. This creates problems in the MC computation procedure which may be partially solved by either using conditional expectation with respect to the noises generated up to τ_N, kernel density estimation methods or integration by parts formulas.

We will denote the transition densities from x to y associated with the forward and backward methods respectively by $q_t^F(x,y)$ and $q_t^B(x,y)$. That is,

$$q_t^F(x,y) = \varphi_{a(x)t}\big(y - x - b(x)t\big), \qquad q_t^B(x,y) = \varphi_{a(x)t}\big(y - x + b(x)t\big).$$

For statements that are true for both the forward and backward methods, we will simply write $q_t(x,y)$.

We note that for the backward method a formula better suited for simulation is obtained by conditioning on all the noise up to τ_N in Equation (5). That is,

$$\mathsf{E}\big[f(X_T)\big] = e^{\lambda T}\mathsf{E}\left[q^B_{T-\tau_N}\big(X^\pi_{\tau_N}, X_0\big)\prod_{i=0}^{N-1}\lambda^{-1}\theta_{\tau_{i+1}-\tau_i}\big(X^\pi_{\tau_i}, X^\pi_{\tau_{i+1}}\big)\right]. \qquad (6)$$

We will henceforth refer to the MC simulation method based on (6) as the backward method with exponential sampling.

Now, we can see the connection to the MLMC method of the previous section. Comparing the first line of (2) with (4), we see that both methods sum over the number of discretization steps so that (4) is in some sense also a MLMC method but where the differences of levels is replaced by the weight function θ and finally we integrate over all possible time partitions at each level.

Since the right-hand side of (5) can be sampled, this gives us an unbiased simulation method. Since the time-steps $\tau_{i+1} - \tau_i$ are exponentially distributed we shall refer to this as the forward/backward-method with exponential (time) sampling. With the method being unbiased, the only source of error is the statistical error and we shall therefore in the following sections investigate the variance of the method. The computational work is governed by the probabilities of N which are parametrized by λ.

We also remark here that the above two methods should be considered as examples of possible parametrix methods. In fact, many other types of basic approximations may be used in order to build an expansion similar to (5) (see, e.g., [8]). In this sense, the above expansion is a Taylor like expansion where one can choose a parameter (the so-called parametrix) in order to obtain the expansion. One of the advantages of the parametrix is that it does not require the existence of strong solutions of the s.d.e. This is in comparison to the MLMC which implicitly demands the existence and uniqueness of strong solutions while the original problem of the calculation of $\mathsf{E}[f(X_T)]$ only requires weak existence of solutions to the s.d.e. (1) in order to make sense.

4. Bounds on θ

In order to analyze the variance of the proposed simulation method (3), we need to find bounds on the weight function $\theta_t(x,y)$.

We will use the Gaussian inequalities and the constants $C_{a,p}(\alpha)$ and $C'_{a,p}(\alpha)$ appearing in Lemma A.1 as well as the inequality

$$\varphi_{ta}(y) \le (2\rho_a)^{d/2}\varphi_{2t\bar{a}}(y).$$

Here a is any invertible matrix such that $0 < \underline{a}I \le a \le \overline{a}I$.

Lemma 4.1. *Assume that there exist $\underline{a}, \overline{a} \in \mathbb{R}$ such that $0 < \underline{a}I \le a(x) \le \overline{a}I$, $a, b \in C_b^\alpha(\mathbb{R}^d)$. In the forward method, we further assume $\sigma_j \in C_b^2(\mathbb{R}^d)$, $b \in C_b^1(\mathbb{R}^d)$. Then for $t \in (0, T]$, there exists a constant $C_T > 0$ which depends on T, $\underline{a}$, $\overline{a}$, and the corresponding norms of a and b in each case such that*

$$\left|\theta_t(x, y)^p q_t(x, y)\right| \le \frac{C_T}{t^{p(1-\zeta/2)}} \varphi_{4\bar{a}t}(y - x), \tag{7}$$

where, in general, for the forward method $\zeta = 1$, for the backward method $\zeta = \alpha$. However if a is constant, then in the forward case $\zeta = 2$ and in the backward case $\zeta = 1 + \alpha$.

An explicit expression for the constant C_T in the above result can be found in the proof below. The statement of this lemma will be important in what follows and therefore we will refer to the parameters and hypotheses in each of the four cases above. That is:

Case 1 (Forward case: general). $\sigma_j \in C_b^2(\mathbb{R}^d)$, $b \in C_b^1(\mathbb{R}^d)$. $\zeta = 1$.
Case 2 (Backward case: general). $\sigma_j, b \in C_b^\alpha(\mathbb{R}^d)$. $\zeta = \alpha$.
Case 3 (Forward case: a constant). $b \in C_b^1(\mathbb{R}^d)$. $\zeta = 2$.
Case 4 (Backward: a constant). $b \in C_b^\alpha(\mathbb{R}^d)$. $\zeta = 1 + \alpha$.

Proof of Lemma 4.1. In case 1,

$$\begin{aligned}
&\left|\kappa_t^{i,j}(x, y) q_t^F(x, y)^{1/p}\right| \\
&\quad = \big|\big(\partial_{i,j}^2 a^{i,j}(y) + \partial_i a^{i,j}(y) H_{ta(x)}^j\big(y - x - b(x)t\big) + \partial_j a^{i,j}(y) H_{ta(x)}^i\big(y - x - b(x)t\big) \\
&\qquad + \big(a^{i,j}(y) - a^{i,j}(x)\big) H_{ta(x)}^{i,j}\big(y - x - b(x)t\big)\big)\varphi_{ta(x)}\big(y - x - b(x)t\big)^{1/p}\big| \\
&\quad \le \left((2\rho_a)^{d/(2p)}\|a\|_2 + 2\|a\|_1 \frac{C_{a,p}(1)}{t^{1/2}} + \|a\|_1\|b\|_0 C'_{a,p}(0) + \|a\|_1 \frac{C'_{a,p}(1)}{t^{1/2}}\right) \\
&\qquad \times \varphi_{2t\bar{a}}\big(y - x - b(x)t\big)^{1/p} \\
&\quad = \left(\frac{\|a\|_1}{t^{1/2}}\big(2C_{a,p}(1) + C'_{a,p}(1)\big) + (2\rho_a)^{d/(2p)}\|a\|_2 + \|a\|_1\|b\|_0 C'_{a,p}(0)\right) \\
&\qquad \times \varphi_{2t\bar{a}}\big(y - x - b(x)t\big)^{1/p},
\end{aligned}$$

and

$$\begin{aligned}
&\left|\rho_t^i(x, y) q_t^F(x, y)^{1/p}\right| \\
&\quad = \big|\big\{\partial_i b^i(y) + \big(b^i(y) - b^i(x)\big) H_{ta(x)}^i\big(y - x - b(x)t\big)\big\}\big| \\
&\qquad \times \varphi_{ta(x)}\big(y - x - b(x)t\big)^{1/p} \\
&\quad \le \big(\|b\|_1 (2\rho_a)^{d/(2p)} + \|b\|_1 C_{a,p}(2) + \|b\|_1\|b\|_0 C_{a,p}(1) t^{1/2}\big)\varphi_{2t\bar{a}}\big(y - x - b(x)t\big)^{1/p}.
\end{aligned}$$

Thus, by Lemma A.1(iii),

$$\begin{aligned}\left|\theta_t(x,y)^p q_t^F(x,y)\right| &= \left|\left(\frac{1}{2}\sum_{i,j}\kappa_t^{i,j}(x,y) - \sum_i \rho_t^i(x,y)\right)\varphi_{a(x)t}\big(y-x-b(x)t\big)^{1/p}\right|^p \\ &\le \frac{C_T}{t^{p(1-\zeta/2)}}\varphi_{4\bar{a}t}(y-x).\end{aligned}$$

Here

$$\begin{aligned}C_T := 2^{d/2}e^{1/4\|b\|_0 T a^{-1}}\Bigg[&\frac{d^2}{2}\|a\|_1\big(2C_{a,p}(1)+C'_{a,p}(1)\big) + Td\|b\|_0\|b\|_1 C_{a,p}(1) \\ &+ T^{1/2}d\Bigg(\frac{d((2\rho_a)^{d/(2p)}\|a\|_2 + \|a\|_1\|b\|_0 C'_{a,p}(0))}{2} \\ &+ \|b\|_1\big((2\rho_a)^{d/(2p)} + C_{a,p}(2)\big)\Bigg)\Bigg]^p.\end{aligned}$$

In case 2, with a_H and b_H being the Hölder constants of a and b, similar calculations give

$$\begin{aligned}&\left|\theta_t(x,y)^p q_t^B(x,y)\right| \\ &\quad\le \varphi_{2t\bar{a}}\big(y-x+b(y)t\big) \\ &\qquad\times\left[\frac{d^2a_H}{2}\frac{C'_{a,p}(\alpha)}{t^{1-\alpha/2}} + \frac{d^2a_H}{2}\|b\|_0^\alpha\frac{C'_{a,p}(0)}{t^{1-\alpha}} + db_H\frac{C_{a,p}(\alpha+1)}{t^{(1-\alpha)/2}} + db_H\|b\|_0^\alpha\frac{C_{a,p}(1)}{t^{1/2-\alpha}}\right]^p \\ &\quad\le \frac{C_T}{t^{p(1-\zeta/2)}}\varphi_{4\bar{a}t}(y-x),\end{aligned}$$

where now

$$\begin{aligned}C_T := 2^{d/2}e^{1/4\|b\|_0 T a^{-1}}\Bigg[&\frac{d^2a_H}{2}C'_{a,p}(\alpha) + \frac{d^2a_H}{2}\|b\|_0^\alpha C'_{a,p}(0)T^{\alpha/2} \\ &+ db_H C_{a,p}(\alpha+1)T^{1/2} + db_H\|b\|_0^\alpha C_{a,p}(1)T^{(1+\alpha)/2}\Bigg]^p.\end{aligned}$$

In cases 3 and 4, that is with constant a, all the above calculations have to be repeated in a similar way to give the claimed result. □

As a corollary of Lemma 4.1, we have the following result.

Corollary 4.2. *For $n \in \mathbb{N}$,*

$$\mathsf{E}\left[\left|\Phi\big(X_T^\pi\big)\prod_{j=0}^{n-1}\theta_{s_{j+1}-s_j}\big(X_{s_j}^\pi, X_{s_{j+1}}^\pi\big)\right|^p\right] \le C_T(T-s_n)^{-q}\prod_{j=0}^{n-1}C_T(s_{j+1}-s_j)^{-p(1-\zeta/2)},$$

where $q = 0$ in the forward case and $q = (p-1)d/2$ in the backward case.

Proof. We will do the proof for the backward case. The forward case is similar and left for the reader. First,

$$q_t^B(x,y)^p \le C_T t^{-(p-1)d/2} \varphi_{2t\bar{a}/p}(x-y), \tag{8}$$

where we used Lemma A.1(iii) and direct calculation on the Gaussian density. Then,

$$\begin{aligned}
&\mathsf{E}\Bigg[\Bigg|\Phi\big(X_T^\pi\big)\prod_{j=0}^{n-1}\theta_{s_{j+1}-s_j}\big(X_{s_j}^\pi, X_{s_{j+1}}^\pi\big)\Bigg|^p\Bigg] \\
&\quad = \int f(x_{s_0})\,dx_{s_0}\prod_{j=0}^{n-1}\big\{\big|\theta_{s_{j+1}-s_j}(x_{s_j},x_{s_{j+1}})\big|^p q_{s_{j+1}-s_j}^B(x_{s_j},x_{s_{j+1}})\,dx_{s_{j+1}}\big\} q_{T-s_n}^B(x_{s_n},X_0)^p \\
&\quad \le C_T\int\prod_{j=0}^{n-1}\big\{C_T(s_{j+1}-s_j)^{-p(1-\zeta/2)}\varphi_{4\bar{a}(s_{j+1}-s_j)}(x_{s_{j+1}}-x_{s_j})\,dx_{s_j}\big\} \\
&\qquad \times (T-s_n)^{-(p-1)d/2}\varphi_{2(s_{j+1}-s_j)\bar{a}/p}(x_{s_n}-X_0)\,dx_{s_n} \\
&\quad = C_T(T-s_n)^{-(p-1)d/2}\prod_{j=0}^{n-1}C_T(s_{j+1}-s_j)^{-p(1-\zeta/2)}.
\end{aligned}$$

□

5. The forward simulation method with exponential time sampling

Let us begin by recalling a basic mathematical result which describes the behavior of the forward simulation method.

Theorem 5.1 ([2]). *Assume that $\sigma_j \in C_b^2(\mathbb{R}^d)$, $b \in C_b^1(\mathbb{R}^d)$, $j \in \{1,\dots,d\}$ and that there exist $\underline{a}, \bar{a} \in \mathbb{R}$ such that $0 < \underline{a}I \le a(x) \le \bar{a}I$. Also, assume that $f \in C_c^\infty(\mathbb{R}^d)$. Then the right side of* (4) *for* (F) *converges absolutely at least at the rate $\frac{C_T^n T^{n/2}}{[n/2]!}$ for some positive constant C_T and therefore the equality* (4) *holds.*

Remark 5.2. The assumption $f \in C_c^\infty(\mathbb{R}^d)$ limits the usefulness of the theorem in financial applications where it is common to consider non-differentiable functions, for example, in option pricing. However it is possible to relax this assumption by using classical limiting arguments. We do not address this technical issue here. In fact, in the present setting one can prove that the density of the process X_T exists and it has Gaussian upper bounds (see, for example, [8]). Therefore, one can consider functions f that have non-compact support and sub-Gaussian upper bounds.

In the case of non-bounded coefficient functions, one can perform some smooth bounded approximation of the coefficients in order to apply the parametrix method. In general, applying the parametrix directly, introduces large variations in the method. For more details, see [19].

While Theorem 5.1 guarantees that the simulation method will converge to the correct value, in order to achieve a statistical error of the order $M^{-1/2}$, where M is the number of MC sample paths used for the simulation, we need the variance to be finite. The following two results show that this is not always the case.

Lemma 5.3. *In addition to the assumptions in Theorem 5.1, assume that $a(x) \equiv a > 0$. Then the forward method with exponential sampling has finite variance.*

Proof. By applying Corollary 4.2, with $p = 2$, $\zeta = 2$ and $q = 0$ we get that the variance of the method based on (5) has as leading constant term,

$$\begin{aligned} e^{2\lambda T}\mathsf{E}\left[f\left(X_T^{\pi}\right)^2\prod_{i=0}^{N-1}\lambda^{-2}\theta^2_{\tau_{i+1}-\tau_i}\left(X^{\pi}_{\tau_i}, X^{\pi}_{\tau_{i+1}}\right)\right] \\ = e^{2\lambda T}\sum_{n=0}^{\infty}\lambda^{-2n}\int_{S^n}\mathsf{E}\left[f\left(X_T^{\pi}\right)^2\prod_{j=0}^{n-1}\theta^2_{s_{j+1}-s_j}\left(X^{\pi}_{s_j}, X^{\pi}_{s_{j+1}}\right)\right]ds \qquad (9) \\ \leq C_T e^{2\lambda T}\sum_{n=0}^{\infty}\frac{\lambda^{-2n}(C_T T)^n}{n!} = C_T e^{2\lambda T + C_T T/\lambda^2}. \end{aligned}$$

□

The above result thus shows that in the case of a constant diffusion term, the forward method with exponential sampling will have finite variance. For future reference, we note that the bound in (9) can be written in terms of the Mittag–Leffler function, see Appendix B, as

$$C_T e^{2\lambda T}E_{1,1}\left(\lambda^{-2}C_T T\right).$$

A negative result is the following. This issue will be solved in Section 7 by using importance sampling on the jump times.

Lemma 5.4. *There are choices of a, b and f, such that the forward simulation method (5) has infinite variance.*

Proof. Without loss of generality, we assume that $d = m = 1$, $b(x) \equiv 0$ and $a \in C_b^2(\mathbb{R})$ and assume that $a' \neq 0$, ν-a.e. and that f is not zero on a set of positive Lebesgue measure. Then for $\bar{\pi} = \{0, s, T\}$,

$$\begin{aligned} \mathsf{E}\left[f^2\left(X_T^{\pi}\right)\prod_{i=1}^{N}\theta^2_{\tau_{i+1}-\tau_i}\left(X^{\pi}_{\tau_{i-1}}, X^{\pi}_{\tau_i}\right)\right] &\geq \mathsf{E}\left[\mathbb{I}(N=1)f^2\left(X_T^{\pi}\right)\theta^2_{\tau_1}\left(X_0^{\pi}, X^{\pi}_{\tau_1}\right)\right] \\ &= \int_0^T \mathsf{E}\left[f^2\left(X_T^{\bar{\pi}}\right)\theta_s^2\left(X_0^{\bar{\pi}}, X_s^{\bar{\pi}}\right)\right]ds. \end{aligned}$$

Define $X_s^{\bar{\pi}} = X_0^{\bar{\pi}} + \sigma(X_0^{\bar{\pi}})\sqrt{s}Z_1$ and $X_T^{\bar{\pi}} = X_s^{\bar{\pi}} + \sigma(X_s^{\bar{\pi}})\sqrt{T-s}Z_2$. We then get for some $X^* \in [X_0^{\bar{\pi}} \wedge X_s^{\bar{\pi}}, X_0^{\bar{\pi}} \vee X_s^{\bar{\pi}}]$

$$\begin{aligned}
&\theta_s\big(X_0^{\bar{\pi}}, X_s^{\bar{\pi}}\big) \\
&\quad = \frac{1}{2}a''\big(X_s^{\bar{\pi}}\big) - a'\big(X_s^{\bar{\pi}}\big)\frac{X_s^{\bar{\pi}} - X_0^{\bar{\pi}}}{a(X_0^{\bar{\pi}})s} + \frac{1}{2}\big(a\big(X_s^{\bar{\pi}}\big) - a\big(X_0^{\bar{\pi}}\big)\big)\left(\left(\frac{X_s^{\bar{\pi}} - X_0^{\bar{\pi}}}{a(X_0^{\bar{\pi}})s}\right)^2 - \frac{1}{a(X_0^{\bar{\pi}})s}\right) \\
&\quad = \frac{1}{2}a''\big(X_s^{\bar{\pi}}\big) + a'\big(X_s^{\bar{\pi}}\big)\frac{Z_1}{\sigma(X_0^{\bar{\pi}})\sqrt{s}} \\
&\qquad + \frac{1}{2}\left(a'\big(X_s^{\bar{\pi}}\big)\sigma\big(X_0^{\bar{\pi}}\big)\sqrt{s}Z_1 - \frac{a''(X^*)a(X_0^{\bar{\pi}})sZ_1^2}{2}\right)\left(\frac{Z_1^2 - 1}{a(X_0^{\bar{\pi}})s}\right) \\
&\quad = \frac{1}{\sqrt{s}}\frac{a'(X_s^{\bar{\pi}})}{\sigma(X_0^{\bar{\pi}})}\frac{Z_1 + Z_1^3}{2} + \frac{1}{2}a''\big(X_s^{\bar{\pi}}\big) - \frac{a''(X^*)Z_1^2(Z_1^2 - 1)}{4}.
\end{aligned}$$

Now,

$$\begin{aligned}
\liminf_{s\to 0} s\mathsf{E}\big[f^2\big(X_T^{\bar{\pi}}\big)\theta_s^2\big(X_0^{\bar{\pi}}, X_s^{\bar{\pi}}\big)\big] &\geq C \liminf_{s\to 0}\mathsf{E}\big[f^2\big(X_T^{\bar{\pi}}\big)a'\big(X_s^{\bar{\pi}}\big)^2\big(Z_1 + Z_1^3\big)^2\big] \\
&\geq C\mathsf{E}\big[f^2\big(X_0^{\bar{\pi}} + \sigma\big(X_0^{\bar{\pi}}\big)\sqrt{T}Z_2\big)a'\big(X_0^{\bar{\pi}}\big)^2\big(Z_1 + Z_1^3\big)^2\big] \geq C.
\end{aligned}$$

Therefore, we can find $\delta > 0$ such that

$$\int_0^T \mathsf{E}\big[f^2\big(X_T^{\bar{\pi}}\big)\theta_s^2\big(X_0^{\bar{\pi}}, X_s^{\bar{\pi}}\big)\big]\,ds \geq \int_0^\delta \frac{C}{2s}\,ds = \infty. \qquad \square$$

Remark 5.5. We should remark here that the fact that the variance is not finite is more a practical issue than a theoretical problem. In fact, the strong law of large numbers still applies even if the variance does not exist. Therefore, the convergence of the method is still assured. The fact that the variance is infinite implies that the convergence will exhibit large deviations from the expectation. The amount and the height of these oscillations will depend on the behavior of moments of order less than 2 as determined by the Marcinkiewicz–Zygmund strong law. In fact, in the case exhibited above all moments of order less than 2 are finite and therefore the deviations from the mean are somewhat limited. However, from a practical point of view, having finite variance is convenient when obtaining confidence intervals for the estimated values.

6. The backward simulation method with exponential time sampling

The backward simulation method is based on the following result.

Theorem 6.1 ([2]). *Assume that there are $\underline{a}, \overline{a} \in \mathbb{R}$ such that $0 < \underline{a}I \le a(x) \le \overline{a}I$, $a, b \in C_b^\alpha(\mathbb{R}^d)$ and $f \in C_c^\infty(\mathbb{R}^d)$ is a density function. Then* (4) *holds for the backward method where the sum converges absolutely at a rate of at least $\frac{C_T^n T^{n\alpha/2}}{[n(\alpha/2)]!}$ for some positive constant C_T.*

The following result shows that the backward method can be expected to perform poorly in dimensions higher than 1.

Lemma 6.2. *In addition to the assumptions in Theorem* 6.1, *assume that $a(x) \equiv a > 0$ and $b \in C_b^\alpha(\mathbb{R}^d)$. Then the backward method with exponential sampling*, (6), *has finite p moment, where $0 < p < \min\{\frac{2}{1-\alpha}, \frac{2}{d} + 1\}$.*

Proof. We apply Corollary 4.2, with $\zeta = 1 + \alpha$ and $q = (p-1)d/2$ and get,

$$\begin{aligned}
&\mathsf{E}\left[\left|q^B_{T-s_n}\bigl(X^\pi_{s_n}, X_0\bigr)\prod_{j=0}^{n-1}\theta_{s_{j+1}-s_j}\bigl(X^\pi_{s_j}, X^\pi_{s_{j+1}}\bigr)\right|^p\right]\\
&\quad\le C_T(T-s_n)^{-(p-1)d/2}\prod_{j=0}^{n-1}C_T(s_{j+1}-s_j)^{-p(1-\alpha)/2}.
\end{aligned}$$

This is integrable over S_n when $p < \min\{\frac{2}{1-\alpha}, \frac{2}{d} + 1\}$. Thus for the p-moment, we have that

$$\begin{aligned}
&e^{p\lambda T}\mathsf{E}\left[q^B_{T-s_n}\bigl(X^\pi_{s_n}, X_0\bigr)^p\prod_{i=0}^{N-1}\lambda^{-p}\bigl|\theta_{\tau_{i+1}-\tau_i}\bigl(X^\pi_{\tau_i}, X^\pi_{\tau_{i+1}}\bigr)\bigr|^p\right]\\
&\quad\le C_T e^{p\lambda T}\sum_{n=0}^\infty \lambda^{-pn}\int_{S^n} C_T^n(T-s_n)^{-(p-1)d/2}\prod_{j=0}^{n-1}(s_{j+1}-s_j)^{-p(1-\alpha)/2}\,ds\\
&\quad= C_T e^{p\lambda T}T^{-(p-1)d/2}\Gamma\bigl(1-(p-1)d/2\bigr)\\
&\qquad\times E_{1-p(1-\alpha)/2,1-(p-1)d/2}\bigl(\lambda^{-p}C_T T^{1-p(1-\alpha)/2}\Gamma\bigl(1-p(1-\alpha)/2\bigr)\bigr),
\end{aligned} \tag{10}$$

where in the last equality we used the definition of the Mittag–Leffler function in Appendix B. □

In the backward case, we thus get a weaker result than in the forward case. In particular for dimensions 2 or greater, the result does not guarantee that the variance of (B) will be finite. In the important special case $p = 2$ and $d = 1$, (10) simplifies to

$$C_T e^{2\lambda T}T^{-1/2}\Gamma(1/2)E_{\alpha,1/2}\bigl(\lambda^{-2}C_T T^\alpha\Gamma(\alpha)\bigr).$$

We remark that the second condition $p < \frac{2}{d} + 1$ appears due to the variance of $q^B_{T-\tau_N}(X^B_{\tau_N}, X_0)$ in (6) which converges to the Dirac delta distribution as $T - \tau_N \to 0$. This will imply that the variance is finite for $d = 1$. In higher dimensions, one solution to the problem is to approximate this distribution by replacing $T - \tau_N$ by $T - \tau_N + \epsilon$ for some small $\epsilon > 0$. This of course introduces a bias and one would have to find an optimal ϵ that balances variance and bias. This can be done using the well-known kernel density estimation techniques.

If we consider the rate of degeneration of the variance, this problem may be improved in polynomial orders by using the Malliavin–Thalmaier integration by parts formula. In particular, one needs to use the Malliavin–Thalmaier type formula which also implies some kernel density type approximation. This will change the bound $p < \min\{\frac{2}{1-\alpha}, \frac{2}{d} + 1\}$ into $p < 2(1 - \epsilon)$ for any $\epsilon > 0$ and $d > 1$. This method which requires an approximation of the Poisson kernel will also introduce bias in the estimation which is controlled by the value of ϵ.

A solution that retains the unbiasedness is based on importance sampling where the direction of simulation is changed again. Therefore, $q_{T-\tau_N}(X^{\pi}_{\tau_N}, X_0)$ is replaced by $f(X^{\pi}_T)$, for $f \in C_c(\mathbb{R}^d)$ as in Theorem 5.1. We however choose not to treat this problem in more detail here and leave as a possible topic for future research.

We also remark that the variance explosion due to the intermediate time discretization points which appeared in Lemma 5.4 also appears here and it gives as a result the above restriction $p < \frac{2}{1-\alpha}$.

7. Achieving finite variance by importance sampling on discretization time points

In examining the proof of Lemma 5.4, we see that the infinite variance is a consequence of the fact that $\theta^2_s(\cdot,\cdot)$ increases at the rate $1/s$ as $s \to 0$. Conditional on $N = n$ the discretization times used in the non-uniform Euler–Maruyama scheme are distributed as the order statistics of a sequence of n i.i.d. uniformly distributed random variables on $[0, T]$. Then the integral on the last line of the proof of Lemma 5.4 diverges not only in the first level but on all levels.

We aim to change the sampling distribution of the discretization times, thereby moving some of the singularity of θ at $s = 0$ from the integrand to the sampling distribution. An example should help to illustrate the idea.

7.1. Toy example

Consider the problem of calculating $\int_0^1 t^\rho \, dt$, for $\rho > -1$. For us, this serves as a much simplified version of (4). As we did with (4), we may rewrite by exchanging the integral for an expectation as follows

$$\int_0^1 t^\rho \, dt = \mathsf{E}\left[X^\rho\right], \qquad X \sim U(0,1),$$

which can be calculated using simulation of n i.i.d. copies of X^ρ with $X \sim U(0,1)$. The above expression corresponds to (5).

Now, if $\rho \in (-1, -1/2]$, the second moment of the random variable X^ρ is

$$\mathsf{E}\left[X^{2\rho}\right] = \int_0^1 t^{2\rho} \, dt = \infty,$$

and thus our simulation will have an exploding variance. This means that the Monte Carlo simulation will exhibit (high) oscillations although there is almost sure convergence. One solution is to use importance sampling. That is, let $p > 1$ and Y be a random variable with density function $t^{-\gamma}(1-\gamma)$, for $0 < t < 1$ and $-\frac{p\rho+1}{p-1} < \gamma < 1$. We then have

$$\int_0^1 t^\rho \, dt = \int_0^1 \frac{t^{\rho+\gamma}}{(1-\gamma)} \frac{(1-\gamma)}{t^\gamma} \, dt = \frac{1}{1-\gamma} \mathsf{E}\big[Y^{\rho+\gamma}\big].$$

And furthermore, the p-moment of the above random variable is always finite as

$$\begin{aligned} \frac{1}{(1-\gamma)^p} \mathsf{E}\big[Y^{p(\rho+\gamma)}\big] &= \frac{1}{(1-\gamma)^p} \int_0^1 t^{p(\rho+\gamma)} \frac{(1-\gamma)}{t^\gamma} \, dt \\ &= \frac{1}{(1-\gamma)^{p-1}(p\rho+(p-1)\gamma+1)}. \end{aligned}$$

Similarly, consider the problem of calculating the infinite sum $\sum_{n=0}^\infty a_n$, for $a_n \geq 0$ for all n. We could formulate this in a probabilistic way by introducing a probability function $p_n > 0$, $n \geq 0$ and writing

$$\sum_{n=0}^\infty a_n = \sum_{n=0}^\infty p_n \frac{a_n}{p_n} = \mathsf{E}\left[\frac{a_N}{p_N}\right],$$

where the random variable N is distributed according to p_n. It is easily seen that the variance minimizing choice of sampling distribution is $p_n = a_n / \sum_{n=0}^\infty a_n$. Although this choice is in practice not available since we do not know $\sum_{n=0}^\infty a_n$ we may use the general heuristic of sampling those n for which a_n is large.

7.2. Importance sampling of discretization time points

We saw in the previous section that by passing some of the singularity of the random variable of interest at 0 to the sampling density, that is, by using importance sampling, we are able to reduce the variance. The following result sets up the importance sampling that we will use in our simulation method.

Lemma 7.1. *Let $\{p_n(s_1,\dots,s_n)\}_{n\geq 0}$ be a family of strictly positive functions, $p_n : S^n \to \mathbb{R}_+$. Suppose that there exists a discrete non-negative random variable N, such that*

$$\mathbb{P}(N=n) = \int_{S^n} p_n(s_1,\dots,s_n)\, ds > 0, \qquad n \geq 1. \tag{11}$$

Also, suppose that there exists a family $\{\tau_i\}_{i\in\mathbb{N}}$ of strictly increasing positive random variables with density conditional on $N = n$, given by $p_n(s_1,\dots,s_n)/\mathbb{P}(N=n)$. Then, for any $g_n \in L^1(S^n)$, $n = 1,\dots,$ the following probabilistic representation holds

$$\int_{S^n} g_n(s_1,\dots,s_n)\, ds = \mathsf{E}\left[\frac{g_N(\tau_1,\dots,\tau_N)}{p_N(\tau_1,\dots,\tau_N)} 1(N=n)\right].$$

Proof. We have that

$$\begin{aligned}
&\mathsf{E}\left[\frac{g_N(\tau_1,\dots,\tau_N)}{p_N(\tau_1,\dots,\tau_N)}1(N=n)\right]\\
&\quad=\mathbb{P}(N=n)\mathsf{E}\left[\frac{g_n(\tau_1,\dots,\tau_n)}{p_n(\tau_1,\dots,\tau_n)}\middle|N=n\right]\\
&\quad=\mathbb{P}(N=n)\int_{S^n}\frac{g_n(s_1,\dots,s_n)}{p_n(s_1,\dots,s_n)}\frac{p_n(s_1,\dots,s_n)}{\mathbb{P}(N=n)}\,ds\\
&\quad=\int_{S^n}g_n(s_1,\dots,s_n)\,ds.
\end{aligned}$$

□

The functions p_n in the previous lemma can be chosen rather arbitrarily. However, firstly, we wish to apply importance sampling to (5) which involves a product. Secondly, an arbitrary choice of p_n could be hard to sample from. Therefore, we consider multiplicative p_n, corresponding to independent increments.

Lemma 7.2. *Let* $\{\xi_i; i\in\mathbb{N}\}$ *be a sequence of i.i.d. random variables with support on* $[0,T+\varepsilon]$, $\varepsilon>0$, *and common strictly positive density* $f_\xi(x)$, $x\in[0,T+\varepsilon]$. *Also, let* $\tau_0=0$, $\tau_i\equiv\sum_{j=1}^{i}\xi_j$, $i\ge 1$ *and let* $N:=\inf\{n;\tau_n<T\le\tau_{n+1}\}$. *Then,* N, τ_i *and the functions*

$$p_n(s_1,\dots,s_n)=\int_{T-s_n}^{T+\varepsilon}f_\xi(x)\,dx\prod_{i=0}^{n-1}f_\xi(s_{i+1}-s_i),\qquad s_0=0,(s_1,\dots,s_n)\in S^n,$$

satisfy the assumptions in Lemma 7.1.

Proof. First, note that the positivity of p_n is clearly satisfied. Furthermore,

$$\mathbb{P}(N=n|\tau_n=s_n)=\int_{T-s_n}^{T+\varepsilon}f_\xi(x)\,dx,$$

and therefore

$$\mathbb{P}\big(N=n,(\tau_1,\dots,\tau_n)\in A\big)=\int_{A\cap S^n}\mathbb{P}(N=n|\tau_n=s_n)\prod_{i=0}^{n-1}f_\xi(s_{i+1}-s_i)\,ds.$$

In particular, $\int_{S^n}p_n(s_1,\dots,s_n)\,ds=\mathbb{P}(N=n)$. Also, the density of $\tau_1,\tau_2,\dots,\tau_n$ conditioned on $N=n$ is given by

$$\frac{\mathbb{P}(N=n|\tau_n=s_n)\prod_{i=0}^{n-1}f_\xi(s_{i+1}-s_i)}{\mathbb{P}(N=n)}=\frac{p_n(s_1,\dots,s_n)}{\mathbb{P}(N=n)}.$$

□

We may now formulate two explicit examples of importance sampling for the time discretization points. These methods will then have improved moment properties.

Proposition 7.3 (Beta sampling). *Let $\{\xi_j; j \in \mathbb{N}\}$ be a sequence of i.i.d. random variables with common density $f_\xi(x) = \frac{(1-\gamma)}{x^\gamma \bar{\tau}^{1-\gamma}}$, $0 < x < \bar{\tau}$, $\bar{\tau} > T$, $\gamma \in (0,1)$ and let N and τ_i be defined as in Lemma 7.2. Then, under the same assumptions as in Theorem 5.1 for the forward and Theorem 6.1 for the backward, the following representation holds*

$$\mathsf{E}\big[f(X_T)\big] = \mathsf{E}\left[\frac{\Phi(X_T^\pi)}{p_N(\tau_1,\dots,\tau_N)}\prod_{j=0}^{N-1}\theta_{\tau_{j+1}-\tau_j}\big(X_{\tau_j}^\pi, X_{\tau_{j+1}}^\pi\big)\right], \tag{12}$$

with

$$p_n(s_1,\dots,s_n) = \left(1 - \left(\frac{T-s_n}{\bar{\tau}}\right)^{1-\gamma}\right)\left(\frac{1-\gamma}{\bar{\tau}^{1-\gamma}}\right)^n \prod_{i=0}^{n-1}\frac{1}{(s_{i+1}-s_i)^\gamma}, \qquad n \geq 0.$$

Also, for the forward method, the p moment of the r.v. inside the expectation of (12) is finite for $p(1-\frac{\zeta}{2}-\gamma) < 1-\gamma$. In the backward method we additionally need that $p < 2/d + 1$, thus the variance is only finite in dimension 1. The values of ζ are given in Lemma 4.1. In particular, if $1-\zeta < \gamma < 1$ then the variance of the random variable in (12) is finite and if we choose $\gamma = 1 - \frac{\zeta}{2}$ then all moments are finite.

Proof. Set

$$g_n(s_1,\dots,s_n) \equiv \mathsf{E}\left[\Phi\big(X_T^\pi\big)\prod_{j=0}^{n-1}\theta_{s_{j+1}-s_j}\big(X_{s_j}^\pi, X_{s_{j+1}}^\pi\big)\right].$$

Note that under Theorem 5.1 for the forward and Theorem 6.1 for the backward, we have that $\sum_{n=0}^\infty |\int_{S^n} g_n(s_1,\dots,s_n)\,ds| < \infty$ and

$$\mathsf{E}\big[f(X_T)\big] = g_0 + \sum_{n=1}^\infty \int_{S^n} g_n(s_1,\dots,s_n)\,ds.$$

The cumulative distribution function of ξ can be found to be $F_\xi(x) = (\frac{x}{\bar{\tau}})^{1-\gamma}$. Using Lemma 7.2, p_n satisfies the assumption in Lemma 7.1 with

$$\begin{aligned} p_n(s_1,\dots,s_n) &= \mathbb{P}(\xi > T - s_n)\prod_{i=0}^{n-1}\frac{(1-\gamma)}{(s_{i+1}-s_i)^\gamma \bar{\tau}^{1-\gamma}} \\ &= \left(1 - \left(\frac{T-s_n}{\bar{\tau}}\right)^{1-\gamma}\right)\left(\frac{1-\gamma}{\bar{\tau}^{1-\gamma}}\right)^n \prod_{i=0}^{n-1}\frac{1}{(s_{i+1}-s_i)^\gamma}. \end{aligned} \tag{13}$$

Thus,

$$\mathsf{E}\big[f(X_T)\big] = \mathsf{E}\left[\frac{g_N(\tau_1,\dots,\tau_N)}{p_N(\tau_1,\dots,\tau_N)}\right] = \mathsf{E}\left[\frac{\Phi(X_T^\pi)}{p_N(\tau_1,\dots,\tau_N)}\prod_{j=0}^{N-1}\theta_{\tau_{j+1}-\tau_j}\big(X_{\tau_j}^\pi, X_{\tau_{j+1}}^\pi\big)\right],$$

and we also note that

$$p_n(s_1,\dots,s_n) \geq \left(1-\left(\frac{T}{\bar{\tau}}\right)^{1-\gamma}\right)\left(\frac{1-\gamma}{\bar{\tau}^{1-\gamma}}\right)^n \prod_{i=0}^{n-1} \frac{1}{(s_{i+1}-s_i)^\gamma}.$$

Using Corollary 4.2, we get

$$\begin{aligned}
&\mathsf{E}\left[\left|\frac{\Phi(X_T^\pi)}{p_N(\tau_1,\dots,\tau_N)}\prod_{i=0}^{N-1}\theta_{\tau_{i+1}-\tau_i}\left(X_{\tau_i}^\pi, X_{\tau_{i+1}}^\pi\right)\right|^p\right] \\
&= \sum_{n=0}^{\infty}\int_{S^n}\mathsf{E}\left[\left|\frac{\Phi(X_T^\pi)}{p_n(s_1,\dots,s_n)}\prod_{i=0}^{n-1}\theta_{s_{i+1}-s_i}\left(X_{s_i}^\pi, X_{s_{i+1}}^\pi\right)\right|^p\right]p_n(s_1,\dots,s_n)\,ds \\
&\leq C_T\sum_{n=0}^{\infty}\int_{S^n}(T-s_n)^{-q}\prod_{j=0}^{n-1}C_T(s_{j+1}-s_j)^{-p(1-\zeta/2)}\frac{1}{p_n(s_1,\dots,s_n)^{p-1}}\,ds \\
&\leq C_T\sum_{n=0}^{\infty}C_T^n\int_{S^n}(T-s_n)^{-q}\prod_{j=0}^{n-1}(s_{j+1}-s_j)^{-(p(1-\zeta/2)-\gamma(p-1))}\,ds,
\end{aligned} \tag{14}$$

the above quantity is finite if $q<1$ and $p(1-\frac{\zeta}{2}-\gamma)<1-\gamma$. □

Remark 7.4. In the above lemma, $\xi_j \overset{d}{=} \bar{\tau}B$, where B is a random variable with a Beta$(1-\gamma,1)$ distribution. This is equivalent to $\xi_j \overset{d}{=} \bar{\tau}e^{-(1-\gamma)E}$, if E has an $\mathsf{Exp}(1)$ distribution. We could instead consider a Beta$(1-\gamma,1-\tilde{\gamma})$ distribution thereby gaining an extra degree of freedom when choosing parameters. However, our main concern is the singularity close to zero which we control by choosing γ appropriately. Adding a parameter $\tilde{\gamma}$ allows us to shift probability mass to the right, but this can also be achieved by choosing $\bar{\tau}$ large. Also, since the shape of the distribution to the right of T is not important we chose not to include a $\tilde{\gamma}$ in our calculations.

Interpretation of the importance sampling method on discretization time points: Note that in the extreme case that $\gamma=0$ then ξ_j has a $\mathsf{U}(0,\bar{\tau})$ distribution. Choosing a parameter $\gamma>0$ means that the algorithm is likely to take more and smaller time discretization steps on average. It thus means that the algorithm will be sampling farther into the sum in (4). If we consider $p=2$ in (14), we can see this clearly. The integral behaves asymptotically as $C_T^n T^{n(\alpha+\gamma-1)}/\Gamma(1+n(\alpha+\gamma-1))$ when $n\to\infty$, and therefore, C_T, T and $\alpha+\gamma$ will determine how far into the sum we need to sample to get a good estimate.

Similarly, a large $\bar{\tau}$ means that the algorithm is likely to take larger and fewer time discretization steps. We must have $\bar{\tau}>T$ since otherwise the algorithm will never sample the term corresponding to $n=0$ in (4), that is, the case with no intermediate time steps between 0 and T. In many cases, it could be possible to calculate this term exactly, as it is an integral w.r.t. the Gaussian measure. In these cases, we may set $\bar{\tau}=T$, thereby possibly gaining efficiency. In this light, one may also propose other alternative importance sampling methods. As an example, we also briefly discuss the following importance sampling method based on Gamma distributions.

Proposition 7.5 (Gamma sampling). *Let $\{\xi_i; i \in \mathbb{N}\}$ be a sequence of i.i.d. r.v.'s with common Gamma$(1-\gamma, \vartheta)$ distribution. That is their common density function is given by $f_\xi(x) = \frac{1}{\Gamma(1-\gamma)\vartheta^{1-\gamma}} \frac{1}{x^\gamma} e^{-x/\vartheta}$, $x > 0$, $1-\alpha < \gamma < 1$, $\vartheta > 0$ and let N and τ_i be defined as in Lemma 7.2. Then the conclusions in Lemma 7.3 holds with*

$$p_n(s_1, \dots, s_n) = \frac{\Gamma(1-\gamma, (T-s_n)/\vartheta)}{\Gamma(1-\gamma)} \left(\frac{1}{\Gamma(1-\gamma)\vartheta^{1-\gamma}} \right)^n e^{-s_n/\vartheta} \prod_{i=0}^{n-1} \frac{1}{(s_{i+1}-s_i)^\gamma},$$

where

$$\Gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} \, dt,$$

is the upper incomplete gamma function.

Proof. It is enough to note that

$$p_n(s_1, \dots, s_n) \geq \frac{\Gamma(1-\gamma, T/\vartheta)}{\Gamma(1-\gamma)} \left(\frac{1}{\Gamma(1-\gamma)\vartheta^{1-\gamma}} \right)^n e^{-T/\vartheta} \prod_{i=0}^{n-1} \frac{1}{(s_i - s_{i-1})^\gamma},$$

and then the arguments in the proof of Lemma 7.3 applies. □

The parameter ϑ in the Gamma distribution roughly corresponds to the parameter $\bar{\tau}$ in the Beta distribution. However, the Gamma distribution has the advantage that ϑ is allowed to take any positive value while $\bar{\tau} > T$. Thus, the Gamma sampling may have an advantage of being more flexible. On the other hand, while we can use the inverse method to generate Beta random variables, generating Gamma random variables is more complicated.

Another way of interpreting the importance sampling procedure is to see that the procedure above chooses a sampling density for the time steps that is similar to $\theta_t(x, y)$, thereby shifting the singularity of $\theta_t(x, y)$ to the sampling density and therefore reducing the variance. This choice also implies a choice of the distribution of N, that is, which levels of the infinite sum we tend to sample from. As we saw in Section 7.1, we should sample those levels in (4) for which the summand is large. But for which levels the summand is large is determined by $\theta_t(x, y)$ and therefore a good choice of sampling density for the time steps will lead to a good choice of levels for which we are sampling frequently.

We also remark that choosing a distribution for the time steps is equivalent to choosing an intensity function, or hazard rate, for the Poisson process for which the time until the first event is the time step. That is, let $\lambda(t)$ be the time-varying intensity of a Poisson process. Then this relates to the density function of the time until the first event, ξ, as,

$$\lambda(t) = \frac{f_\xi(t)}{1 - F_\xi(t)},$$

$$f_\xi(x) = \lambda(x) e^{-\int_0^x \lambda(t) \, dt}.$$

For example, $\lambda(t) = \frac{(1-\gamma)t^{-\gamma}}{\bar{\tau}^{1-\gamma}-t^{1-\gamma}}$, $t \in (0, \bar{\tau})$, implies a Beta distribution,

$$\lambda(t) = \frac{\beta^{1-\gamma}t^{-\gamma}}{\Gamma(1-\gamma)\Gamma(1-\gamma, \beta t)} e^{-\beta t}, \qquad t > 0,$$

implies a Gamma distribution and $\lambda(t) = \alpha t^{-\gamma}$ $t > 0$, implies a Weibull distribution. We however feel that in the importance sampling context it is more natural to specify the distribution of time steps and not the intensity function.

8. Optimal parameters

After introducing the importance sampling methods proposed in the previous sections, we will now discuss how to choose the parameters of the method in order to maximize the efficiency.

8.1. Complexity and parameter optimization

The complexity of our algorithm depends on the choice of the importance sampling parameters. This is because they affect the number of time steps that the simulation will take on average.

Define the process $N_t = \sup\{n | \sum_{i=1}^{n} \tau_i \le t\} + 1$. $N_t - 1$ is thus a renewal process and $N \equiv N_T$ is the number of time steps in our algorithm. One can expect the running time of the algorithm to be approximately proportional to $\mathsf{E}[N]$ and we thus take this to be the complexity. In general, it is difficult to calculate $\mathsf{E}[N]$ and to the best of our knowledge there are no closed formulas for the case of Beta and Gamma distributed inter-arrival times. In the Gamma case, we may use that the sum of Gamma distributed random variables is again Gamma and we then have that

$$\mathsf{E}[N] = \sum_{n=1}^{\infty} P(N \ge n) = \sum_{n=1}^{\infty} P\left(\sum_{i=1}^{n} \tau_i \le T\right) = \sum_{n=1}^{\infty}\left(1 - \frac{\Gamma(n(1-\gamma), T/\theta)}{\Gamma(n(1-\gamma))}\right).$$

From renewal theory, we however know that $\mathsf{E}[N] < \infty$ and the elementary renewal theorem tells us that $\lim_{t\to\infty} \mathsf{E}[N_t - 1]/t = 1/\mathsf{E}[\tau_i]$. We use it to motivate the following approximations,

$$\begin{aligned} \mathrm{Exp}(\lambda) &: \mathsf{E}[N] = T\lambda + 1, \\ \bar{\tau}\,\mathrm{Beta}(1-\gamma, 1) &: \mathsf{E}[N] \approx \frac{T}{\mathsf{E}[\tau_i]} + 1 = \frac{T}{\bar{\tau}} \frac{2-\gamma}{(1-\gamma)} + 1, \\ \mathrm{Gamma}(1-\gamma, \vartheta) &: \mathsf{E}[N] \approx \frac{T}{\mathsf{E}[\tau_i]} = \frac{T}{\vartheta} \frac{1}{(1-\gamma)} + 1. \end{aligned}$$

Comparing the above with Remark 7.4 , we see that γ small will imply lower complexity of the simulation scheme. On the other hand, values of γ close to 1 imply that we will be examining higher order terms in (4).

Let $V(p)$ be the variance of a single sample from the simulation algorithm which depends on a parameter p for the importance sampling procedure. We will define the efficiency of the algorithm to be the inverse of the product of the computational work and $V(p)$. We will define the computational work to be the average number of time steps in the method, that is, $\mathsf{E}[N]$. That this is a good measure of efficiency is rigorously motivated using limit theorems in [13]. Thus, our optimization problem is

$$\min_p E[N]V(p). \tag{15}$$

In general, it will be difficult to find the exact theoretical value of the quantity $V(p)$. For this reason, we will address a minimization problem for an upper bound of $E[N]V(p)$. It thus remains to find an upper bound of $V(p)$ in order to be able to carry out the minimization procedure.

8.2. Optimal parameters with exponential sampling in the constant diffusion case

The purpose of this section is to give a benchmark in the case where the time sampling is done using the exponential distribution and the diffusion coefficient is constant. This will be used later when comparing with other time sampling schemes. The parameter that we optimize over here is λ, corresponding to p in the previous section.

The bound on the variance in the forward and backward method with exponential sampling of the time steps can be summarized as

$$C_T e^{2\lambda T} T^{-q} \Gamma(1-q) E_{\alpha,1-q}\bigl(\lambda^{-2} C_T T^{\alpha} \Gamma(\alpha)\bigr),$$

where in the forward method $q = 0$ and in the backward $q = 1/2$. Thus, the optimization problem (15) becomes,

$$\min_{\lambda} (\lambda T + 1) e^{2\lambda T} T^{-q} \Gamma(1-q) E_{\alpha,1-q}\bigl(\lambda^{-2} C_T T^{\alpha} \Gamma(\alpha)\bigr).$$

From the definition of the Mittag–Leffler function, we can see that $E_{\alpha,\beta}(z)$ is convex, increasing and non-negative for $z \in \mathbb{R}^+$, thus the above objective function is convex and therefore its minimum exists uniquely and is finite.

After a careful calculation, we may conclude that in the forward case, the optimal λ is increasing in C_T and T. Thus, a s.d.e. with less regular coefficients will require a simulation method that on average uses more time steps. Note that the actual value of C_T can be obtained from the proof of Lemma 5.3 and therefore we see that the constant C_T may be small in particular cases. This seems to be the case in many financial models.

In the backward case, it is more difficult to make the analogous conclusions. However, our numerical results indicate that the above conclusions in the forward method are also valid in the backward method.

8.3. Optimal parameters with Beta importance sampling

In this section, we will derive an upper bound on the variance of our estimator in the case of Beta sampling, the case of Gamma sampling is analogous. Then we will study the minimization problem as in Section 8.1 for this upper bound.

We now again apply Corollary 4.2 and make the parameter change $\beta = \zeta + \gamma - 1$, noting that $0 < \beta < \zeta$. We call β the distance to non-integrability. In fact, ζ measures coefficient regularity, γ measures the importance sampling index and -1 comes from the degeneration of the corresponding Hermite polynomials from θ. We will optimize over β and $\bar{\tau}$, corresponding to the p in (15).

Letting $E_{\alpha,\beta}$ denote the Mittag–Leffler function (see Appendix B), we have for $C_T \frac{\bar{\tau}^{\zeta-\beta}}{\zeta-\beta} T^\beta \Gamma(\beta)$ large enough and $\beta \in (0, 2)$

$$
\begin{aligned}
&\mathsf{E}\left[\left(\frac{\Phi(X_T^\pi)}{p_N(\tau_1, \dots, \tau_N)} \prod_{i=0}^{N-1} \theta_{\tau_{i+1}-\tau_i}\left(X_{\tau_i}^\pi, X_{\tau_{i+1}}^\pi\right)\right)^2\right] \\
&\quad \le C_T \sum_{n=0}^{\infty} \int_{S^n} C_T^n (T - s_n)^{-q} \prod_{i=0}^{n-1} (s_{i+1} - s_i)^{-(2-\zeta)} \frac{1}{p_n(s_1, \dots, s_n)} \, ds \\
&\quad \le C_T \left(1 - \left(\frac{T}{\bar{\tau}}\right)^{\zeta-\beta}\right)^{-1} \\
&\qquad \times \left[\sum_{n=0}^{\infty} C_T^n \left(\frac{\bar{\tau}^{\zeta-\beta}}{\zeta - \beta}\right)^n \int_{S^n} (T - s_n)^{-q} \prod_{i=0}^{n-1} (s_{i+1} - s_i)^{-(1-\beta)} \, ds\right] \qquad (16) \\
&\quad = C_T \left(1 - \left(\frac{T}{\bar{\tau}}\right)^{\zeta-\beta}\right)^{-1} T^{-q} \Gamma(1 - q) E_{\beta,1-q}\left(C_T \frac{\bar{\tau}^{\zeta-\beta}}{\zeta - \beta} T^\beta \Gamma(\beta)\right) \\
&\quad \approx C_T \left(1 - \left(\frac{T}{\bar{\tau}}\right)^{\zeta-\beta}\right)^{-1} \frac{1}{\beta} \Gamma(1 - q) \left(C_T \frac{\bar{\tau}^{\zeta-\beta}}{\zeta - \beta} \Gamma(\beta)\right)^{q/\beta} \exp\left(T \left(C_T \frac{\bar{\tau}^{\zeta-\beta}}{\zeta - \beta} \Gamma(\beta)\right)^{1/\beta}\right) \\
&\quad \equiv V(\beta, \bar{\tau}). \qquad (17)
\end{aligned}
$$

The approximation in the last step is exact for $\beta = 1$, $q = 0$ and performs well when $z = C_T \frac{\bar{\tau}^{\zeta-\beta}}{\zeta-\beta} T^\beta \Gamma(\beta) > 1$ (see Figure 1). However, in situations where this is less than 1, the variance will be relatively small and there is less need to choose simulation parameters optimally. Also note that, in our application $\beta \in (0, 2)$ is fulfilled.

Here we also see that β measures, in some sense, the distance to non-integrability of the integral in (16). That is, for $0 \le \beta < 1$, the integrand of (16) is L^p integrable in $(s_1, \dots, s_{n-1})$ for $p < \frac{1}{1-\beta}$.

We have from (17) that

$$
\lim_{\beta \to 0} \mathsf{E}[N] V(\beta, \bar{\tau}) = \lim_{\beta \to \zeta} \mathsf{E}[N] V(\beta, \bar{\tau}) = \lim_{\bar{\tau} \to T} \mathsf{E}[N] V(\beta, \bar{\tau}) = \lim_{\bar{\tau} \to \infty} \mathsf{E}[N] V(\beta, \bar{\tau}) = \infty.
$$

So, by continuity, $\mathsf{E}[N] V(\beta, \bar{\tau})$ achieves its absolute minimum in $0 < \beta < \zeta$, $T < \bar{\tau} < \infty$.

Defining $F(\bar{\tau}, C_T) \equiv \partial_{\bar{\tau}} \log(\mathsf{E}[N] V(\beta^\star, \bar{\tau}))$, a strict optimal $\bar{\tau}$ solves $F(\bar{\tau}, C_T) = 0$, with $\partial_{\bar{\tau}} F(\bar{\tau}, C_T) \ge 0$. Now, assuming that $\beta^\star$ remains locally constant as a function of C_T (see Fig-

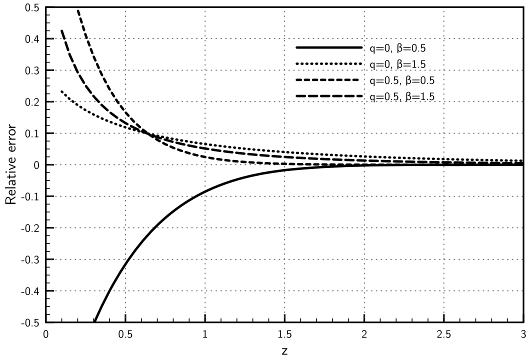


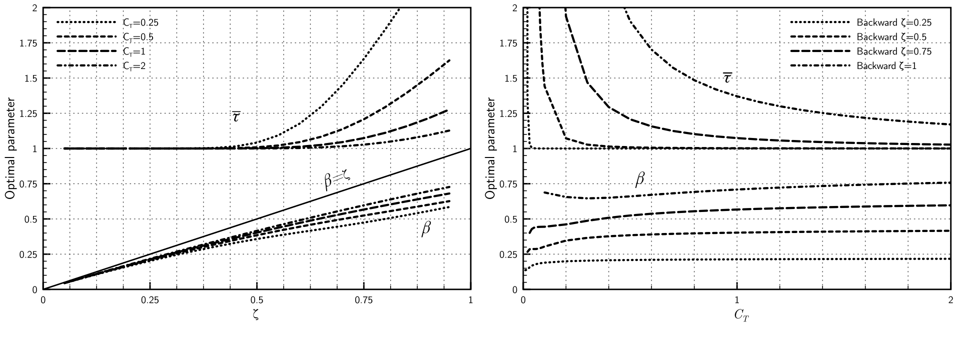
Figure 1. Relative error of exponential approximation of $E_\beta(z)$.

ure 2(a)) since

$$\partial_{C_T} F(\bar{\tau}, C_T) = \frac{T(\zeta - \beta^\star)(C_T \Gamma(\beta^\star)\bar{\tau}^{\zeta-\beta^\star})^{1/\beta^\star}}{C_T {\beta^\star}^2 \bar{\tau}} > 0,$$

we get by implicit differentiation that $\frac{\partial \bar{\tau}}{\partial C_T} \leq 0$. Thus as general rule we should choose $\bar{\tau}$ relatively large, thus sampling the lower levels often, when C_T is small, and vice versa. Note that this is the same heuristic conclusion as in Section 8.2.

As the minimization problem can not be solved exactly, we solve it numerically in some cases and plot the result in Figure 2. We see that as ζ increases, $\bar{\tau}$ increases. Also $\zeta - \beta$ increases, implying that the optimal importance sampling index γ is decreasing. Values of ζ closer to 1 corresponds to Lipschitz regularity of the coefficients of the s.d.e. Thus for a s.d.e. in such a case the algorithm can take larger time steps. In Figure 2(b), we see that $\bar{\tau}$ is decreasing in C_T.



(a) Varying ζ (b) Varying C_T

Figure 2. Optimal parameters in the backward method.

This situation is analogous to the dependence on ζ. Small C_T corresponds to a regular s.d.e., in a different sense than for ζ. Thus, when C_T is small it is possible to take large time steps. The dependence of β on C_T is not as strong. This can be understood if we remember that β has the purpose of taking care of the integrability, which is not affected by C_T. It is also worth noting that although the optimal $\bar{\tau}$ increases fast for large ζ and small C_T the actual difference in the Beta distribution and the variance bounds may not be that large.

In summary, these results confirm the heuristics that a non-regular s.d.e. requires smaller time steps, by choosing a small β and $\bar{\tau}$. We also remark that this is analogous to the conclusions regarding the dependence of the optimal λ on C_T in Section 8.2.

For Gamma sampling, the bound becomes

$$\begin{aligned}
&\mathsf{E}\Biggl[\biggl(\frac{\Phi(X_T^\pi)}{p_N(\tau_1,\dots,\tau_N)}\prod_{i=0}^{N-1}\theta_{\tau_{i+1}-\tau_i}\bigl(X_{\tau_i}^\pi, X_{\tau_{i+1}}^\pi\bigr)\biggr)^2\Biggr] \\
&\quad\le C_T e^{T/\vartheta}\frac{\Gamma(\zeta-\beta)}{\Gamma(\zeta-\beta, T/\vartheta)}T^{-q}\Gamma(1-q)E_{\beta,1-q}\bigl(C_T\vartheta^{\zeta-\beta}T^\beta\Gamma(\zeta-\beta)\Gamma(\beta)\bigr) \\
&\quad\approx C_T\frac{\Gamma(\zeta-\beta)\Gamma(1-q)}{\Gamma(\zeta-\beta, T/\vartheta)}\frac{e^{T/\vartheta}}{\beta}\bigl(C_T\Gamma(\zeta-\beta)\theta^{\zeta-\beta}\Gamma(\beta)\bigr)^{q/\beta} \\
&\qquad\times\exp\bigl(T\bigl(C_T\vartheta^{\alpha-\beta}\Gamma(\zeta-\beta)\Gamma(\beta)\bigr)^{1/\beta}\bigr).
\end{aligned}$$

Note also that as the above minimization results will differ from every actual application. We can only interpret the results as classes. That is, for any class of functions f, a and b such that the constant C_T is smaller than a certain value the above minimization problem result can be applied. In that sense, the stability properties in the figures are of interest.

9. Simulations

In this section, we apply the simulation methods on two test cases. We begin by treating an s.d.e. which is expected to show the difference between sampling the random time steps from an Exponential distribution and a Beta distribution and to confirm our general rules on how to choose the simulation parameters.

Further, we treat a model that shows that in the case of a Hölder continuous diffusion part, choosing γ large enough will give a finite variance and thus a fast rate of convergence.

We note, as in Remark 5.2, that although in the examples below $f(x) \notin C_c^\infty(\mathbb{R}^d)$, the results from the previous sections could be extended to the examples considered here.

9.1. Choosing simulation parameters

We shall in this section consider the solution of the s.d.e.

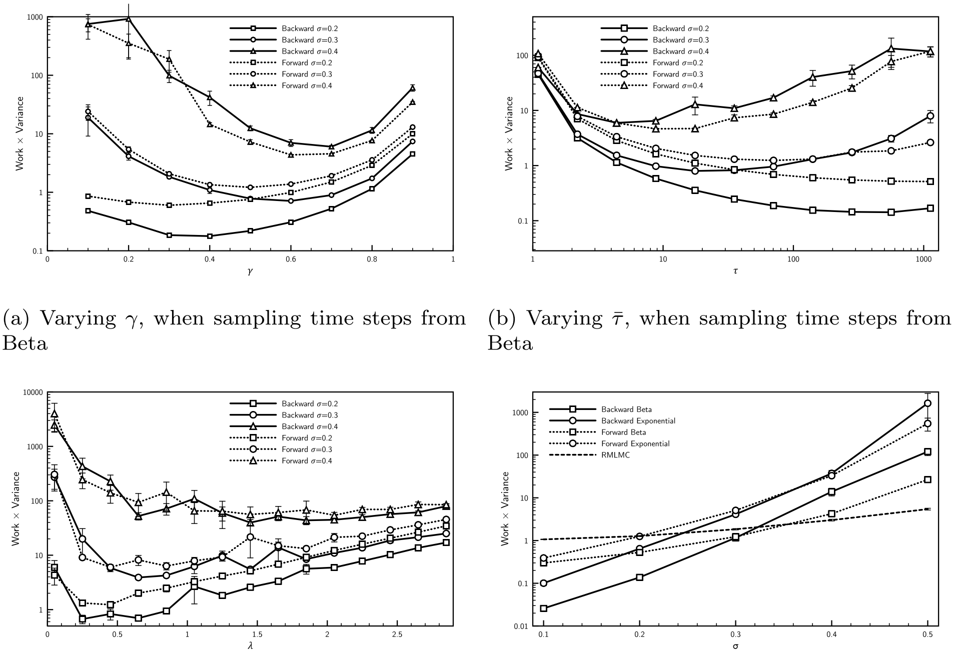
$$dX_t = \sigma\bigl(\sin(\omega X_t) + 2\bigr)\,dW_t,$$

for $\sigma > 0$. Also note that the assumptions of Theorems 5.1 and 6.1 are fulfilled and we thus expect both the forward and backward method to converge. For both the forward and backward method $\zeta = 1$, thus $\beta = \gamma - 1$.

We choose to simulate $P(X_0 - I < X_T < X_0 + I)$, where I is such that the probability is approximately 0.5, respectively for each parameter choice. Throughout, $T = 1$ and $X_0 = 0$.

We examine how the performance of the forward and backward methods depends on the choice of simulation parameters, that is, in Exponential time sampling with parameter λ and in Beta time sampling with parameters γ and $\bar{\tau}$. We measure the performance of the method using work × variance, where we measure work as the total number of time steps used in the algorithm. In terms of the optimization problem (15), $\mathsf{E}[N]$ is replaced by the actual work needed to achieve the variance $V(p)$. We should note here that since the variance is not finite in the case of Exponential time sampling, the sample variance is not a good measure of performance. We include it here however as a comparison to the performance obtained using Beta sampling.

We use three different parameter sets, $\sigma = \omega = 0.2$, 0.3 and 0.4. The results are given in Figure 3. We see that for γ, $\bar{\tau}$ and λ the curve appears convex and so there is an optimal choice. All three parameters also appear to follow the general heuristics. That is, the parameter $\sigma = \omega$ becomes larger, it is necessary to take smaller time steps, i.e. choose larger γ and λ or smaller $\bar{\tau}$.



(a) Varying γ, when sampling time steps from Beta

(b) Varying $\bar{\tau}$, when sampling time steps from Beta

(c) Varying λ, when sampling time steps from Exp

(d) Using optimal simulation parameters and varying $\sigma = \omega$

Figure 3. Work × variance from simulations. Error bars represent 1 standard deviation. (Some error bars are too small to be seen.)

Although not illustrated in the figures, it should also be mentioned here that since γ and $\bar{\tau}$ have opposite effects on the distribution of the time steps, that is, increasing γ or decreasing $\bar{\tau}$ gives smaller time steps and vice versa, an increase in γ can somewhat be canceled by a decrease in $\bar{\tau}$. Thus, the optimal value of one parameter will depend on the choice of the other.

In Figure 3(d), we see the performance of the different methods, for close to optimal choices of simulation parameters, as $\sigma = \omega$ varies. Most notably, the performance of the method deteriorates quickly as σ increases. In fact, for larger values, the variance becomes difficult to estimate from the simulations and the estimates become unreliable. We also see that the Beta sampling method outperforms the Exponential sampling method, at least for larger σ, while it is difficult to draw any general conclusions about the difference in performance between the backward and the forward method.

Also in Figure 3(d), we compare the forward and backward method to the unbiased randomized multilevel Monte Carlo method (RMLMC) described in [17]. We have implemented what is referred to in [17] as the single-term estimator and we find the optimal sampling distribution by estimating the variance of the 10 first terms, using 10^4 samples, and assuming geometrically declining variances after that. Note that to implement the RMLMC one needs to know the strong order of convergence of the Milstein scheme, which is used in the method. However in our case, where f is not Lipschitz, the order is not known and so it is not clear exactly how the implement the method. We however implement it as if f was Lipschitz, with the understanding that the theorems in [17] do not apply. There is however a method described in [9], where the pay-off function is smoothed using conditional expectations, which could be applied here.

We see that the parametrix methods seem to outperform the RMLMC for smaller σ will the opposite seems to hold for larger σ. Of course, these results may depend on the particular implementation of the methods.

9.2. Convergence rate

In this section, we consider

$$dX_t = k(X_0 - X_t)\,dt + \sigma\sqrt{|X_0 - X_t|^{1/4} + 1}\,dW_t,$$

where $k = 1.5$, $X_0 = 1$, $\sigma = 0.01$. As in the previous section we wish to simulate $P(X_0 - I < X_T < X_0 + I)$, where I is such that the probability is approximately 0.5 and $T = 1$.

We first note that the diffusion part is not differentiable so that the forward method is not applicable. Also, the drift and diffusion is not bounded, so Theorem 6.1 does not apply. Nonetheless, this is an interesting case since the diffusion is only Hölder continuous and our results in Section 7 suggests that choosing γ large enough in the Beta sampling should produce a finite variance. Second, the process is mean-reverting towards the point where the diffusion is Hölder. This is necessary since to see the effect of the Hölder continuity we need the process to visit this point frequently. We should also remark that this is also the reason that we have chosen σ quite small.

We use the Backward method with Beta sampling for $\gamma = 0.1$ and 0.9. For increasing sample sizes, we calculate the mean absolute error. We also calculate the rate of convergence as the slope

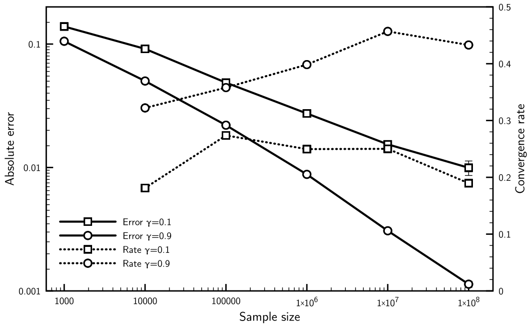


Figure 4. Absolute error and convergence rate with for different γ. Error bars for the absolute error represent 1 standard deviation. (Most error bars are too small to be seen.)

of the error on the log-scale. The result can be seen in Figure 4. We see that the larger γ gives a smaller absolute error for all sample sizes and more importantly that the rate of convergence is faster. While theory implies that in the finite variance case, we should get a convergence rate of 0.5, we get a slightly slower rate at the largest sample size. We believe this to be a finite sample size effect. That is, that if we were able to make the sample size larger we would approach the rate 0.5. On the other hand, the smaller γ gives a convergence rate well below 0.5, indicating that the variance is not finite.

10. Conclusion

The main goal of the present paper is to analyze the performance of the simulation method which stems from (4). We found that the forward method works well in particular cases. For example, if the diffusion coefficient is a constant matrix then the variance of the method is finite. In other cases the variance may be infinite and the simulation method will then suffer from a poor convergence rate. In these situations, we propose an importance sampling method on the time steps, using a beta or gamma distribution. This will improve the performance of the method if the parameters of the importance sampling method are chosen correctly.

As it is usually done in MLMC, we also study the minimization of variance given a restriction on computational time or vice-versa. This gives us the tools needed to find good parameters for the importance sampling distribution. We also find certain heuristic guidelines, such as that irregular s.d.e.'s need parameters that takes more and therefore smaller time steps. Thus, the importance sampling distribution needs to have more mass closer to 0.

Finally, we provide some simulations to demonstrate the performance of the method. The simulations confirm the theoretical findings. We also find that the simulation method works well when parameters are not too large. For larger parameters, the variance becomes large. This problem may be solved using some importance sampling methods also on the space variables.

There are many issues that have to be studied in the future. In particular, to study by simulation higher dimensional examples, the implementation of a deterministic time partition in order to reduce variance, space importance sampling methods, parametrix methods based on fixed discrete time grids and applications to various other stochastic equations remain as some of the subjects to be studied. This gives a glimpse of the flexibility and the applicability of the method.

Still the problem of the explosion of variance remains an important issue. One solution is proposed here. There are various other possibilities that one may also entertain if one is willing to accept again some bias in the method. Such is the case of the localization of weight functions θ between others.

Appendix A: Gaussian inequalities

In order to explicitly state the bounds for the variances, we define the constant $C_{a,p}(\alpha) := (2\rho_a)^{d/(2p)}(4\bar{a}p)^{\alpha/2}\underline{a}^{-1}$.

Lemma A.1. *For* $\alpha \in [0,1]$, $p>0$, $y \in \mathbb{R}^d$, $a \in \mathbb{R}^{d\times d}$ *such that* $0 < \underline{a}I_{d\times d} \le a \le \bar{a}I_{d\times d}$ *and* $t>0$:

(i)

$$|y|^\alpha \left|H^i_{ta}(y)\right|\varphi_{ta}(y)^{1/p} \le \frac{C_{a,p}(\alpha+1)}{t^{(1-\alpha)/2}}\varphi_{2t\bar{a}}(y)^{1/p}.$$

(ii) *Define* $C'_{a,p}(\alpha) := C_{a,p}(2+\alpha)\underline{a}^{-1} + C_{a,p}(\alpha)$ *then*

$$|y|^\alpha \left|H^{ij}_{ta}(y)\right|\varphi_{ta}(y)^{1/p} \le \frac{C'_{a,p}(\alpha)}{t^{1-\alpha/2}}\varphi_{2t\bar{a}}(y)^{1/p}.$$

(iii) *There exists a constant* $C_T = 2^{d/2}e^{(1/4)\|b\|_0 T a^{-1}}$ *such that*

$$\varphi_{2ta}\bigl(y-x-b(x)t\bigr) \le C_T\varphi_{4ta}(y-x).$$

Proof. First, note that

$$\left|H^i_{ta}(y)\right| \le \frac{|y|}{t\underline{a}}, \qquad \left|H^{i,j}_{ta}(y)\right| \le \frac{|y|^2}{t^2\underline{a}^2} + \frac{1}{t\underline{a}},$$
$$\varphi_{ta}(y) \le \rho_a^{d/2}\varphi_{t\bar{a}}(y) = (2\rho_a)^{d/2}\exp\left\{-\frac{1}{4t\bar{a}}|y|^2\right\}\varphi_{2t\bar{a}}(y).$$

For the proof of (i) we have that,

$$|y|^\alpha\left|H^i_{ta}(y)\right|\varphi_{ta}(y)^{1/p} \le \frac{|y|^{\alpha+1}}{t\underline{a}}(2\rho_a)^{d/(2p)}\exp\left\{-\frac{1}{4t\bar{a}p}|y|^2\right\}\varphi_{2t\bar{a}}(y)^{1/p}.$$

We shall also use that $v^r e^{-v} \le 1$ for $v \ge 0$ and $0 \le r \le 1$. Here, take $v = \frac{1}{4t\bar{a}p}|y|^2$ and $r = \frac{\alpha+1}{2}$ and the inequality follows.

For the proof of inequality (ii), we have that

$$|y|^\alpha \left|H^{ij}_{ta}(y)\right| \varphi_{ta}(y)^{1/p} \le \frac{|y|^\alpha}{t\underline{a}}\left(\frac{|y|^2}{t\underline{a}}+1\right)(2\rho_a)^{d/(2p)} \exp\left\{-\frac{1}{4t\overline{a}}|y|^2\right\}\varphi_{2t\overline{a}}(y)^{1/p}.$$

Now, repeating the same argument as in the proof of (i) with $v = \frac{|y|^2}{4t\overline{a}}$, $r = \alpha/2$ and $r = \frac{2+\alpha}{2}$, we get (ii).

The proof of (iii) follows by direct calculation. In fact, using Young's inequality $|(y-x)b(x)| \le \frac{|y-x|^2}{2} + \frac{|b(x)|^2}{2}$, we obtain the result. □

Appendix B: Mittag–Leffler functions

We need that for $\rho, \eta < 1$,

$$\begin{aligned}
&\sum_{n=0}^{\infty} C^n \int_{S^n} (T-s_n)^{-\eta} \prod_{i=0}^{n-1} (s_{i+1}-s_i)^{-\rho}\, ds \\
&\quad = T^{-\eta}\Gamma(1-\eta) \times \sum_{n=0}^{\infty} C^n T^{n(1-\rho)} \frac{\Gamma^n(1-\rho)}{\Gamma(1-\eta+n(1-\rho))} \\
&\quad = T^{-\eta}\Gamma(1-\eta) E_{1-\rho,1-\eta}\bigl(CT^{1-\rho}\Gamma(1-\rho)\bigr),
\end{aligned}$$

where

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\beta+\alpha k)}, \qquad z \in C, \alpha, \beta > 0,$$

is the Mittag–Leffler function, see, for example, [7]. Some special cases are

$$\begin{aligned}
E_{0,1}(z) &= \frac{1}{1-z}, \qquad |z| < 1, \\
E_{1/2,1}\bigl(\pm z^{1/2}\bigr) &= e^z \operatorname{erfc}\bigl(\mp z^{1/2}\bigr), \\
E_{1,1}(z) &= e^z.
\end{aligned}$$

We also have that

$$E_{\alpha,\beta}(z) = \alpha^{-1} z^{(1-\beta)/\alpha} \exp\bigl(z^{1/\alpha}\bigr) + O\bigl(|z|^{-1}\bigr), \qquad 0 < \alpha < 2, |\arg z| < \pi/2, |z| \to \infty.$$

Later we will use the approximation $E_{\alpha,\beta}(z) \approx \alpha^{-1} z^{(1-\beta)/\alpha} \exp(z^{1/\alpha})$, somewhat abusing the above limit approximation.

Acknowledgment

The research by Patrik Andersson was funded by the Swedish Research Council and done while visiting Ritsumeikan University.

References

[1] Avikainen, R. (2009). On irregular functionals of SDEs and the Euler scheme. *Finance Stoch*. **13** 381–401. MR2519837

[2] Bally, V. and Kohatsu-Higa, A. (2015). A probabilistic interpretation of the parametrix method. *Ann. Appl. Probab*. **25** 3095–3138. MR3404632

[3] Bally, V. and Talay, D. (1996). The law of the Euler scheme for stochastic differential equations. II. Convergence rate of the density. *Monte Carlo Methods Appl*. **2** 93–128. MR1401964

[4] Beskos, A., Papaspiliopoulos, O. and Roberts, G.O. (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli* **12** 1077–1098. MR2274855

[5] Broadie, M. and Kaya, Ö. (2006). Exact simulation of stochastic volatility and other affine jump diffusion processes. *Oper. Res*. **54** 217–231. MR2222897

[6] Chen, B., Oosterlee, C.W. and van der Weide, H. (2012). A low-bias simulation scheme for the SABR stochastic volatility model. *Int. J. Theor. Appl. Finance* **15** 1250016, 37. MR2911733

[7] Erdélyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F.G. (1955). *Higher Transcendental Functions. Vol. III*. New York: McGraw-Hill Book Company, Inc. MR0066496

[8] Friedman, A. (1964). *Partial Differential Equations of Parabolic Type*. Englewood Cliffs, NJ: Prentice-Hall, Inc. MR0181836

[9] Giles, M. (2008). Improved multilevel Monte Carlo convergence using the Milstein scheme. In *Monte Carlo and Quasi-Monte Carlo Methods* 2006 343–358. Berlin: Springer. MR2479233

[10] Giles, M.B. (2008). Multilevel Monte Carlo path simulation. *Oper. Res*. **56** 607–617. MR2436856

[11] Giles, M.B. and Szpruch, L. (2013). Antithetic multilevel Monte Carlo estimation for multidimensional SDEs. In *Monte Carlo and Quasi-Monte Carlo Methods* 2012. *Springer Proc. Math. Stat*. **65** 367–384. Heidelberg: Springer. MR3145572

[12] Glynn, P.W. (1983). Randomized estimators for time integrals. Technical report, Mathematics Research Center, University of Wisconsin, Madison.

[13] Glynn, P.W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Oper. Res*. **40** 505–520. MR1180030

[14] Gyöngy, I. and Rásonyi, M. (2011). A note on Euler approximations for SDEs with Hölder continuous diffusion coefficients. *Stochastic Process. Appl*. **121** 2189–2200. MR2822773

[15] McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods Appl*. **17** 301–315. MR2890424

[16] Rhee, C. and Glynn, P.W. (2012). A new approach to unbiased estimation for SDE's. In *Simulation Conference* (*WSC*), *Proceedings of the* 2012 *Winter* 1–7. Berlin: IEEE.

[17] Rhee, C. and Glynn, P.W. (2015). Unbiased estimation with square root convergence for s.d.e. models. *Oper. Res*. **63** 1026–1043.

[18] Stroock, D.W. and Varadhan, S.R.S. (2006). *Multidimensional Diffusion Processes*. *Classics in Mathematics*. Berlin: Springer. MR2190038

[19] Tanaka, Y. (2015). On the approximation of the stochastic differential equation via parametrix method. Master's thesis, Graduate School of Science and Engineering, Ritsumeikan University, Japan.

Received September 2014 and revised September 2015