

PRACTICAL VARIANCE REDUCTION VIA REGRESSION FOR SIMULATING DIFFUSIONS*

G. N. MILSTEIN[†] AND M. V. TRETYAKOV[‡]

Abstract. The well-known variance reduction methods—the method of importance sampling and the method of control variates—can be exploited if an approximation of the required solution is known. Here we employ conditional probabilistic representations of solutions together with the regression method to obtain sufficiently inexpensive (although rather rough) estimates of the solution and its derivatives by using the single auxiliary set of approximate trajectories starting from the initial position. These estimates can effectively be used for significant reduction of variance and further accurate evaluation of the required solution. The developed approach is supported by numerical experiments.

Key words. probabilistic representations of solutions of partial differential equations, numerical integration of stochastic differential equations, Monte Carlo technique, variance reduction methods, regression

AMS subject classifications. Primary, 65C05; Secondary, 65C30, 60H10

DOI. 10.1137/060674661

1. Introduction. The stochastic approach to solving problems of mathematical physics is based on probabilistic representations of their solutions by making use of the weak-sense numerical integration of stochastic differential equations (SDEs) and the Monte Carlo (MC) technique. In this approach we have two main errors: the error of SDE numerical integration and the MC error. The first error essentially depends on the choice of a method of numerical integration, and the second one depends on the choice of the probabilistic representation (it is understood that the first error for a chosen method can be reduced by decreasing the step of discretization, and the MC error for a selected probabilistic representation can be reduced by increasing the number of independent trajectories). While the error of numerical integration is well studied in the systematic theory of numerical integration of SDEs, which allows us to propose suitable effective methods for a lot of typical problems (see, e.g., [16]), in connection with the MC error there is a lack of constructive variance reduction methods.

The well-known variance reduction methods (see [12, 16, 21] and the references therein) of importance sampling and of control variates can be exploited only in the case when an approximation of the required solution $u(t, x)$ is known. However, in general even rough approximations of the desired solution $u(t, x)$ and its derivatives $\partial u / \partial x^i(t, x)$, $i = 1, \dots, d$, are unknown beforehand. At first sight, it seems that approximating them roughly is not difficult since they can be found by the MC technique using a comparatively small number of independent trajectories. But this presupposes evaluating them at many points (t_k, x_k) . Computing $u(t_k, x_k)$ and $\partial u / \partial x^i(t_k, x_k)$ by

*Received by the editors November 10, 2006; accepted for publication (in revised form) October 17, 2008; published electronically February 6, 2009. This work was partially supported by the Royal Society International Joint Project-2004/R2-FS grant and UK EPSRC research grant EP/D049792/1. <http://www.siam.org/journals/sinum/47-2/67466.html>

[†]Department of Mathematics, Ural State University, Lenin Str. 51, 620083 Ekaterinburg, Russia (Grigori.Milstein@usu.ru).

[‡]Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK (M.Tretyakov@le.ac.uk). This author's research was partially supported by a Leverhulme Research Fellowship. Part of this work was done while the author was on study leave granted by the University of Leicester.

the MC technique requires different auxiliary sets of approximate trajectories because of the different starting points (t_k, x_k) . This is too expensive, i.e., as a rule, such a procedure is more expensive than simple increase of the number of trajectories starting from the initial position (t_0, x_0) , at which we aim to find the value of the solution u .

So, a suitable method of constructing $u(t_k, x_k)$ and $\partial u / \partial x^i(t_k, x_k)$ should be comparatively inexpensive. Therefore we cannot require a considerable accuracy of the estimates for $u(t_k, x_k)$ and $\partial u / \partial x^i(t_k, x_k)$ because there is a trade-off between accuracy and computational expenses. Our proposition is to exploit conditional probabilistic representations. Their employment together with the regression method allows us to evaluate $u(t_k, x)$ and $\partial u / \partial x^i(t_k, x)$ using the single auxiliary set of approximate trajectories starting from the initial position (t_0, x_0) only. This plays a crucial role in obtaining sufficiently inexpensive (but at the same time useful for variance reduction) estimates $\hat{u}(t_k, x)$ and $\widehat{\partial u} / \partial x^i(t_k, x)$. The construction of \hat{u} and $\widehat{\partial u} / \partial x^i$ is accompanied by a number of errors of a different nature. Although it is impossible to evaluate these errors satisfactorily, the suitability of $\hat{u}(t_k, x)$ and $\widehat{\partial u} / \partial x^i(t_k, x)$ for variance reduction can be directly verified during computations since the MC error can always be estimated. We emphasize that the obtained (even rather rough) estimates can effectively be used for accurately evaluating the function u not only at the position (t_0, x_0) but at many other positions as well.

This paper is most closely connected with [6, 12, 13, 14] (see also the [16]) and with the works [21, 20] by N. Newton. The method of importance sampling from [6, 12] is exploited in [25] for some specific physical applications. Various other aspects of variance reduction related to simulating diffusions are considered, e.g., in [2, 4, 9, 10, 24] (see also the references therein). An extended list of works devoted to variance reduction of MC simulations can be found in [7].

In section 2 we recall some known facts concerning the MC technique for linear parabolic equations and the general scheme of regression method for estimating conditional expectations. Section 3 is devoted to conditional probabilistic representations of solutions of parabolic equations and their derivatives. These representations together with regression approach play a decisive role in the economical estimating of u and $\partial u / \partial x^i$ at all points (t, x) , given the only set of trajectories starting from the initial point (t_0, x_0) . In section 3.2 we obtain the estimate $\hat{u}(s, x)$ and propose to estimate the derivatives $\partial u / \partial x^i(s, x)$ by $\partial \hat{u} / \partial x^i(s, x)$. This estimation of derivatives is inexpensive from the computational point of view, but they are rather rough. Section 3.3 is devoted to the more accurate way of estimating derivatives using a linear regression method directly to find $\widehat{\partial u} / \partial x^i(t_k, x)$. In section 3.4, we obtain $\widehat{\partial u} / \partial x^i(t_k, x)$ in the case of nonsmooth initial data exploiting probabilistic representations for $\partial u / \partial x^i(s, x)$ which rest on the Malliavin integration by parts. To this aim, we derive a conditional version of the Malliavin integration-by-parts formula adapted to our context. It should be noted that if the dimension d is large, the procedures of sections 3.3 and 3.4 are computationally very demanding since they require integration of the d^2 -dimensional system of first-order variation equations whose solution is present in the probabilistic representations for $\partial u / \partial x^i(s, x)$. Therefore, in practice, the inexpensive procedure of section 3.2 is preferable if d is large. In section 4 we give a simple, analytically tractable example to illustrate the benefits of the proposed variance reduction procedure, and we also test it on a one-dimensional array of stochastic oscillators and on the Black–Scholes pricing model for a binary asset-or-nothing call option. Section 5 gives a summary of the proposed approach to variance reduction.

2. Preliminaries. In this section we recall some known facts concerning probabilistic representations of the solutions of parabolic partial differential equations and the regression method of estimating conditional expectations in the form suitable for our purposes.

2.1. Probabilistic representations. Let us consider the Cauchy problem for the linear parabolic equation

$$(2.1) \quad \frac{\partial u}{\partial t} + \frac{1}{2} \sum_{i,j=1}^d a^{ij}(t, x) \frac{\partial^2 u}{\partial x^i \partial x^j} + \sum_{i=1}^d b^i(t, x) \frac{\partial u}{\partial x^i} + c(t, x)u + g(t, x) = 0, \quad t_0 \leq t < T, \quad x \in \mathbf{R}^d,$$

with the initial condition

$$(2.2) \quad u(T, x) = f(x), \quad x \in \mathbf{R}^d.$$

The matrix $a(t, x) = \{a^{ij}(t, x)\}$ in (2.1) is symmetric and at least positive semidefinite. Let $\sigma(t, x)$ be a matrix obtained from the equation

$$a(t, x) = \sigma(t, x)\sigma^\top(t, x).$$

Let $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, $t_0 \leq t \leq T$, be a filtered probability space. The solution to the problem (2.1)–(2.2) has the following probabilistic representation (the well-known Feynman–Kac formula):

$$(2.3) \quad u(s, x) = E[f(X_{s,x}(T))Y_{s,x,1}(T) + Z_{s,x,1,0}(T)],$$

where $X_{s,x}(t)$, $Y_{s,x,y}(t)$, $Z_{s,x,y,z}(t)$, $t \geq s$, is the solution of the Cauchy problem for the system of SDEs

$$(2.4) \quad \begin{aligned} dX &= b(t, X)dt + \sigma(t, X)dw(t), \quad X(s) = x, \\ dY &= c(t, X)Ydt, \quad Y(s) = y, \\ dZ &= g(t, X)Ydt, \quad Z(s) = z. \end{aligned}$$

Here $w(t) = (w^1(t), \dots, w^d(t))^\top$ is a d -dimensional $\{\mathcal{F}_t\}_{t \geq t_0}$ -adapted standard Wiener process, and Y and Z are scalars. If $y = 1$, $z = 0$, we shall use the notation $Y_{s,x}(t) := Y_{s,x,1}(t)$, $Z_{s,x}(t) := Z_{s,x,1,0}(t)$ (analogous notation will be used later for some other variables). So,

$$(2.5) \quad u(s, x) = E[f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T)].$$

There are various sets of sufficient conditions ensuring connection between the solutions of the Cauchy problem (2.1)–(2.2) and their probabilistic representations (2.5)–(2.4). For definiteness, we shall keep the following assumptions.

We assume that the coefficients b , σ , c , and g have bounded derivatives up to some order, and additionally c and g are bounded on $[t_0, T] \times \mathbf{R}^d$. Further, we assume that the matrix $a(t, x)$ is positive definite and, moreover, the uniform ellipticity condition holds: there exists $\sigma_0 > 0$ such that

$$\|a^{-1}(t, x)\| = \|(\sigma(t, x)\sigma^\top(t, x))^{-1}\| \leq \sigma_0^{-1}, \quad t_0 \leq t \leq T, \quad x \in \mathbf{R}^d.$$

As for function $f(x)$, it is assumed to grow at infinity not faster than a polynomial function. It can be both smooth and nonsmooth.

We note that the results of this paper can be used under other sets of conditions. For instance, one can consider situations with nonglobally Lipschitz coefficients [18] or with matrix $a(t, x)$ which is positive semidefinite. For example, in section 4.2 we consider a numerical example with nonglobally Lipschitz coefficients and positive semidefinite matrix $a(t, x)$, and the example from section 4.3 has a discontinuous $f(x)$.

The value $u(s, x)$ from (2.5) can be evaluated using the weak-sense numerical integration of the system (2.4) together with the MC technique. More specifically, we have

$$(2.6) \quad \begin{aligned} u(s, x) &\approx E[f(\bar{X}_{s,x}(T))\bar{Y}_{s,x}(T) + \bar{Z}_{s,x}(T)] \\ &\approx \frac{1}{M} \sum_{m=1}^M [f({}_m\bar{X}_{s,x}(T)){}_m\bar{Y}_{s,x}(T) + {}_m\bar{Z}_{s,x}(T)], \end{aligned}$$

where the first approximate equality involves an error due to replacing X, Y, Z by $\bar{X}, \bar{Y}, \bar{Z}$ (the error is related to the approximate integration of (2.4)) and the error in the second approximate equality comes from the MC technique; ${}_m\bar{X}_{s,x}(T), {}_m\bar{Y}_{s,x}(T), {}_m\bar{Z}_{s,x}(T), m = 1, \dots, M$, are independent realizations of $\bar{X}_{s,x}(T), \bar{Y}_{s,x}(T), \bar{Z}_{s,x}(T)$. While the weak-sense integration of SDEs is developed sufficiently well and a lot of different effective weak-sense numerical methods have been constructed (see, e.g., [16]), the methods of reducing the second error in (2.6) are more intricate.

The error of the MC method is evaluated by

$$\bar{\rho} = c \frac{(\text{var}[f(\bar{X}_{s,x}(T))\bar{Y}_{s,x}(T) + \bar{Z}_{s,x}(T)])^{1/2}}{M^{1/2}},$$

where, e.g., the values $c = 1, 2, 3$ correspond to the fiducial probabilities 0.68, 0.95, 0.997, respectively. Introduce

$$(2.7) \quad \Gamma = \Gamma_{s,x} := f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T),$$

$$(2.8) \quad \bar{\Gamma} = \bar{\Gamma}_{s,x} := f(\bar{X}_{s,x}(T))\bar{Y}_{s,x}(T) + \bar{Z}_{s,x}(T).$$

Since $\text{var}\Gamma_{s,x}$ is close to $\text{var}\bar{\Gamma}_{s,x}$, we can assume that the error of the MC method is estimated by

$$(2.9) \quad \rho = c \frac{(\text{var}\Gamma_{s,x})^{1/2}}{M^{1/2}}.$$

2.2. Variance reduction. If $\text{var}\Gamma_{s,x}$ is large, then to achieve a satisfactory accuracy we have to simulate a very large number of independent trajectories. Clearly, variance reduction is of crucial importance for effectiveness of any MC procedure. To reduce the MC error, one usually exploits some other probabilistic representations of solutions to considered problems. To obtain various probabilistic representations of the solution to the problem (2.1)–(2.2), we introduce the system (see [13, 14, 16])

$$(2.10) \quad \begin{aligned} dX &= b(t, X)dt - \sigma(t, X)\mu(t, X)dt + \sigma(t, X)dw(t), \quad X(s) = x, \\ dY &= c(t, X)Ydt + \mu^\top(t, X)Ydw(t), \quad Y(s) = 1, \\ dZ &= g(t, X)Ydt + F^\top(t, X)Ydw(t), \quad Z(s) = 0, \end{aligned}$$

where μ and F are column-vector functions of dimension d satisfying some regularity conditions (e.g., they have bounded derivatives with respect to x^i up to some order). We should note that X, Y, Z in (2.10) differ from X, Y, Z in (2.4); however, this does not lead to any ambiguity. The formula (2.5), i.e.,

$$(2.11) \quad u(s, x) = E\Gamma_{s,x},$$

remains valid under the new X, Y, Z . While the mean $E\Gamma$ does not depend on the choice of μ and F , the variance $\text{var}\Gamma = E\Gamma^2 - (E\Gamma)^2$ does. Thus, μ and F can be used to decrease the variance $\text{var}\Gamma$ and, consequently, the MC error can be reduced. The following theorem is proved in [14] (see also [13, 16]).

THEOREM 2.1. *Let μ and F be such that for any $x \in \mathbf{R}^d$ there exists a solution to the system (2.10) on the interval $[s, T]$. Then the variance $\text{var}\Gamma$ is equal to*

$$(2.12) \quad \text{var}\Gamma = E \int_s^T Y_{s,x}^2(t) \sum_{j=1}^d \left(\sum_{i=1}^d \sigma^{ij} \frac{\partial u}{\partial x^i} + u\mu^j + F^j \right)^2 dt,$$

provided that the expectation in (2.12) exists. In (2.12) all the functions $\sigma^{ij}, \mu^j, F^j, u, \partial u/\partial x^i$ have $(t, X_{s,x}(t))$ as their argument.

In particular, if μ and F are such that

$$(2.13) \quad \sum_{i=1}^d \sigma^{ij} \frac{\partial u}{\partial x^i} + u\mu^j + F^j = 0, \quad j = 1, \dots, d,$$

then $\text{var}\Gamma = 0$, i.e., Γ is deterministic.

We recall that if we put here $F = 0$, then we obtain the method of importance sampling (first considered in [6, 12, 24]), and if we put $\mu = 0$, then we obtain the method of control variates (first considered in [21]). Theorem 2.1 establishes the combining method of variance reduction proved in [13]; see also [16].

Obviously, μ and F satisfying (2.13) cannot be constructed without knowing $u(t, x)$, $s \leq t \leq T$, $x \in \mathbf{R}^d$. Nevertheless, the theorem claims a general possibility of variance reduction by a proper choice of the functions μ^j and F^j , $j = 1, \dots, d$. Theorem 2.1 can be used, for example, if we know a function $\hat{u}(t, x)$ connected with an approximating problem and which is close to $u(t, x)$. In this case we take any $\hat{\mu}^j, \hat{F}^j$, $j = 1, \dots, d$, satisfying

$$(2.14) \quad \sum_{i=1}^d \sigma^{ij} \frac{\partial \hat{u}}{\partial x^i} + \hat{u}\hat{\mu}^j + \hat{F}^j = 0,$$

and then the variance $\text{var}\Gamma$, though not zero, is small.

Let us emphasize that (2.13) serves only as a guidance for getting suitable μ and F (recall that the mean $E\Gamma$ does not depend on the choice of μ and F). In particular, the derivative estimate $\widehat{\partial u/\partial x^i}$ can differ from $\partial \hat{u}/\partial x^i$. In such cases, instead of (2.14) we use

$$(2.15) \quad \sum_{i=1}^d \sigma^{ij} \frac{\widehat{\partial u}}{\partial x^i} + \hat{u}\hat{\mu}^j + \hat{F}^j = 0.$$

It might seem that the problem of at least rough approximation of the functions $u(t, x)$ and $\partial u/\partial x^i(t, x)$ is not difficult since they can be found approximately due to

the Feynman–Kac formula, numerical integration of SDEs, and the MC technique. But then numerical integration of the system (2.10) presupposes evaluating $u(t_k, \bar{X}_k)$ and $\partial u / \partial x^i(t_k, \bar{X}_k)$ at many points (t_k, \bar{X}_k) . Their evaluation by the MC method requires different sets of auxiliary approximate trajectories because of the different starting points (t_k, \bar{X}_k) . This is too expensive; i.e., as a rule, such a procedure is more expensive than simple increase of M in (2.6).

Our aim is to propose a systematic method of approximating the functions u and $\partial u / \partial x^i$, $i = 1, \dots, d$, relatively cheaply, and hence obtain systematic methods of variance reduction. To this end, we exploit the regression method of evaluating $u(t_k, x)$ and $\partial u / \partial x^i(t_k, x)$, which allows us to use only one set of approximate trajectories starting from the initial position (t_0, x_0) .

2.3. Pathwise approach for derivatives $\partial u / \partial x^i(s, x)$. The probabilistic representation for the derivatives

$$\partial^i(s, x) := \frac{\partial u(s, x)}{\partial x^i}, \quad i = 1, \dots, d,$$

can be obtained by the straightforward differentiation of (2.11) (see, e.g., [7, 13]):

$$(2.16) \quad \partial^i(s, x) = E \left(\sum_{j=1}^d \frac{\partial f(X_{s,x}(T))}{\partial x^j} \delta_{s,x}^i X^j(T) Y_{s,x}(T) + f(X_{s,x}(T)) \delta_{s,x}^i Y(T) + \delta_{s,x}^i Z(T) \right),$$

where

$$\begin{aligned} \delta^i X^j(t) &:= \delta_{s,x}^i X^j(t) := \frac{\partial X_{s,x}^j(t)}{\partial x^i}, \quad \delta^i Y(t) := \delta_{s,x}^i Y(t) := \frac{\partial Y_{s,x}(t)}{\partial x^i}, \\ \delta^i Z(t) &:= \delta_{s,x}^i Z(t) := \frac{\partial Z_{s,x}(t)}{\partial x^i}, \quad s \leq t \leq T, \quad i, j = 1, \dots, d, \end{aligned}$$

satisfy the system of variational equations associated with (2.10):

$$(2.17) \quad d\delta^i X = \sum_{j=1}^d \frac{\partial(b(t, X) - \sigma(t, X)\mu(t, X))}{\partial x^j} \delta^i X^j dt + \sum_{j=1}^d \frac{\partial\sigma(t, X)}{\partial x^j} \delta^i X^j dw(t),$$

$$\delta^i X^j(s) = 0 \text{ if } j \neq i, \text{ and } \delta^i X^i(s) = 1,$$

$$(2.18) \quad \begin{aligned} d\delta^i Y &= \sum_{j=1}^d Y \frac{\partial c(t, X)}{\partial x^j} \delta^i X^j dt + c(t, X) \delta^i Y dt \\ &+ \sum_{j=1}^d Y \frac{\partial \mu^\top(t, X)}{\partial x^j} \delta^i X^j dw(t) + \mu^\top(t, X) \delta^i Y dw(t), \quad \delta^i Y(s) = 0, \end{aligned}$$

$$(2.19) \quad \begin{aligned} d\delta^i Z &= \sum_{j=1}^d Y \frac{\partial g(t, X)}{\partial x^j} \delta^i X^j dt + g(t, X) \delta^i Y dt \\ &+ \sum_{j=1}^d Y \frac{\partial F^\top(t, X)}{\partial x^j} \delta^i X^j dw(t) + F^\top(t, X) \delta^i Y dw(t), \quad \delta^i Z(s) = 0. \end{aligned}$$

Introduce a partition of the time interval $[t_0, T]$, for simplicity the equidistant one: $t_0 < t_1 < \dots < t_N = T$ with step size $h = (T - t_0)/N$. Let us apply a weak scheme (see, e.g., [16]) to the systems of SDEs (2.10), (2.17)–(2.19) to obtain independent approximate trajectories $(t_k, {}_m\bar{X}(t_k))$, $m = 1, \dots, M$, all starting from the point (t_0, x) , and ${}_m\bar{Y}(t_k)$, ${}_m\bar{Z}(t_k)$, ${}_m\bar{\delta^i X}(t_k)$, ${}_m\bar{\delta^i Y}(t_k)$, ${}_m\bar{\delta^i Z}(t_k)$ with ${}_m\bar{Y}(t_0) = 1$, ${}_m\bar{Z}(t_0) = 0$, ${}_m\bar{\delta^i X^j}(t_0) = 0$ if $j \neq i$, and ${}_m\bar{\delta^i X^i}(t_0) = 1$, ${}_m\bar{\delta^i Y}(t_0) = 0$, ${}_m\bar{\delta^i Z}(t_0) = 0$. Then we obtain the following MC estimates of the derivatives $\partial u / \partial x^i(t_0, x)$ from (2.16) with $(s, x) = (t_0, x)$:

$$(2.20) \quad \hat{\partial}^i(t_0, x) = \frac{1}{M} \sum_{m=1}^M \left[\sum_{j=1}^d \frac{\partial f(m\bar{X}(T))}{\partial x^j} {}_m\bar{\delta}^i X^j(T) {}_m\bar{Y}(T) + f(m\bar{X}(T)) {}_m\bar{\delta}^i Y(T) + {}_m\bar{\delta}^i Z(T) \right].$$

Clearly, the estimates $\hat{\partial}^i(t_k, x)$ for derivatives $\partial u / \partial x^i(t_k, x)$ can be obtained analogously.

Theorem 2.1 asserts that the variance in evaluating u by (2.11) can reach zero value for some μ and F . In [13] it is proved that for the same μ and F the variance in evaluating ∂^2 by (2.16) is equal to zero as well (we pay attention that not only μ and F but also their derivatives are present in (2.18) and (2.19)).

2.4. Regression method of estimating conditional expectation. Let us recall the general scheme of the linear regression method (see, e.g., [8]). Consider a sample $({}_mX, {}_mV)$, $m = 1, \dots, M_r$, from a generic member (X, V) of the sample, where X is a d -dimensional and V is a one-dimensional random variable. We pay attention that we denote by M_r the size of the sample used in the regression, while M is the number of realizations used for computing the required quantity $u(t_0, x_0)$ (see (2.6)). Let the values of X belong to a domain $\mathbf{D} \subset \mathbf{R}^d$. It is of interest to estimate the regression function

$$(2.21) \quad c(x) = E(V|X = x).$$

Let $\{\varphi_l(x)\}_{l=1}^L$ be a set of basis functions each mapping \mathbf{D} to \mathbf{R} . As an estimate $\hat{c}(x)$ of $c(x)$, we choose the function of the form $\sum_{l=1}^L \alpha_l \varphi_l(x)$ that minimizes the empirical risk:

$$(2.22) \quad \hat{\alpha} = \arg \min_{\alpha \in \mathbf{R}^L} \frac{1}{M_r} \sum_{m=1}^{M_r} \left({}_mV - \sum_{l=1}^L \alpha_l \varphi_l({}_mX) \right)^2.$$

So

$$(2.23) \quad \hat{c}(x) = \sum_{l=1}^L \hat{\alpha}_l \varphi_l(x),$$

where $\hat{\alpha}_l$ satisfy the system of linear algebraic equations

[illegible]

with

$$(2.25) \quad a_{ln} = \frac{1}{M_r} \sum_{m=1}^{M_r} \varphi_l(mX) \varphi_n(mX), \quad b_l = \frac{1}{M_r} \sum_{m=1}^{M_r} \varphi_l(mX) {}_mV, \quad l, n = 1, \dots, L.$$

Thus, the usual base material in the field of regression is a sample $({}_mX, {}_mV)$, $m = 1, \dots, M_r$, from a generic member (X, V) of the sample.

Remark 2.2. Although in this paper we use linear regression, in principle other regression methods (see, e.g., [3, 8]) can be exploited as well.

3. Conditional probabilistic representations and methods of evaluating $u(s, x)$ and $\partial u / \partial x^i(s, x)$ by regression. The routine (unconditional) probabilistic representations are ideal for the MC evaluation of $u(t_0, x_0)$ by using a set of trajectories starting from the point (t_0, x_0) . To find $u(s, x)$ by this approach, we need to construct another set of trajectories which starts from (s, x) . However, we can use the previous set starting from (t_0, x_0) to compute $u(s, x)$, $s > t_0$, if we make use of conditional probabilistic representations. In this section we introduce the conditional probabilistic representations for solutions of parabolic equations and for derivatives of the solutions.

3.1. Conditional probabilistic representations for $u(s, x)$ and $\partial u / \partial x^i$ (s, x). Along with the unconditional probabilistic representation (2.11), (2.7), (2.10) for $u(s, x)$, we have the following conditional one:

$$(3.1) \quad \begin{aligned} u(s, x) &= E(f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T)) \\ &= E(f(X_{s,X}(T))Y_{s,X}(T) + Z_{s,X}(T) \text{ with } X := X_{t_0,x_0}(s) | X_{t_0,x_0}(s) = x). \end{aligned}$$

This formula can be considered as the conditional version of the Feynman–Kac formula.

Analogously to (3.1), we get for $\partial^i(s, x) = \partial u / \partial x^i(s, x)$ (see (2.16))

$$(3.2) \quad \begin{aligned} \partial^i(s, x) &= E \left(\sum_{j=1}^d \frac{\partial f(X_{s,x}(T))}{\partial x^j} \delta_{s,x}^i X^j(T) Y_{s,x}(T) + f(X_{s,x}(T)) \delta_{s,x}^i Y(T) + \delta_{s,x}^i Z(T) \right) \\ &= E \left(\sum_{j=1}^d \frac{\partial f(X_{s,X}(T))}{\partial x^j} \delta_{s,X}^i X^j(T) Y_{s,X}(T) \right. \\ &\quad \left. + f(X_{s,X}(T)) \delta_{s,X}^i Y(T) + \delta_{s,X}^i Z(T) | X := X_{t_0,x_0}(s) = x \right). \end{aligned}$$

So, we have two different probabilistic representations both for $u(s, x)$ and $\partial^i(s, x)$: the first one is in the form of unconditional expectation (see section 2), and the second one (i.e., (3.1) and (3.2)) is in the form of conditional expectation. The first form can be realized naturally by the MC approach and the second one by a regression method. As we discussed before, it is too expensive to run sets of trajectories starting from various initial points (s, x) , and we do have the set of trajectories $(t, {}_mX_{t_0,x_0}(t))$. Taking this into account, the second way (which relies on the conditional probabilistic representations and regression) is more preferable although it is less accurate.

A proof of (3.1) and (3.2) relies on the following assertion: if ζ is $\tilde{\mathcal{F}}$ -measurable, $f(x, \omega)$ is independent of $\tilde{\mathcal{F}}$, and $E f(x, \omega) = \phi(x)$, then $E(f(\zeta, \omega) | \tilde{\mathcal{F}}) = \phi(\zeta)$ (see,

e.g., [11]). From this assertion, for any measurable g it holds (with $\zeta = X_{t_0, x_0}(s)$, $\tilde{\mathcal{F}} = \sigma\{X_{t_0, x_0}(s)\}$, $f(x, \omega) = g(X_{s, x}(T))$) that

$$E(g(X_{s, X}(T)) | X_{t_0, x_0}(s) = x) = Eg(X_{s, x}(T)) \text{ with } X := X_{t_0, x_0}(s),$$

hence (3.1) and (3.2).

3.2. Evaluating $u(s, x)$. In evaluating $u(s, x)$ by regression, the pairs (X, V) and $({}_mX, {}_mV)$ have the form

$$(3.3) \quad \begin{aligned} (X, V) &\sim (X_{t_0, x_0}(s), f(X_{s, X}(T))Y_{s, X}(T) + Z_{s, X}(T)), \\ ({}_mX, {}_mV) &\sim ({}_mX_{t_0, x_0}(s), f({}_mX_{s, {}_mX}(T)){}_mY_{s, {}_mX}(T) + {}_mZ_{s, {}_mX}(T)). \end{aligned}$$

To realize a regression algorithm, we construct the set of trajectories $(t, {}_mX_{t_0, x_0}(t))$. Of course, we construct them approximately at the time moments $s = t_k$ and store the obtained values. So, in reality we have $(t_k, {}_m\bar{X}_{t_0, x_0}(t_k))$. The time s in (3.3) is equal to that of t_k . We note that

$$(3.4) \quad X_{s, X}(t) = X_{s, X_{t_0, x_0}(s)}(t) = X_{t_0, x_0}(t), \quad t \geq s;$$

i.e., $X_{s, X}(t)$ is a continuation of the base solution starting at the moment t_0 and $X_{s, X}(T)$ in (3.3) is equal to $X_{t_0, x_0}(T)$. It is not so for Y :

$$Y_{s, X}(T) \neq Y_{t_0, x_0}(T).$$

Let us recall that $Y_{s, X}(t)$ is the solution of the equation (see (2.10))

$$(3.5) \quad dY_{s, X} = c(t, X_{s, X}(t))Y_{s, X}dt + \mu^\top(t, X_{s, X}(t))Y_{s, X}dw(t), \quad Y(s) = 1.$$

Clearly,

$$(3.6) \quad Y_{s, X}(t) = \frac{Y_{t_0, x_0}(t)}{Y_{t_0, x_0}(s)}, \quad s \leq t \leq T,$$

hence storing $Y_{t_0, x_0}(t)$, we can get $Y_{s, X}(T)$ in (3.3).

Analogously, $Z_{s, X}(T) \neq Z_{t_0, x_0}(T)$. It is not difficult to find that

$$(3.7) \quad Z_{s, X}(t) = \frac{1}{Y_{t_0, x_0}(s)}(Z_{t_0, x_0}(t) - Z_{t_0, x_0}(s)), \quad Z_{s, X}(T) = \frac{1}{Y_{t_0, x_0}(s)}(Z_{t_0, x_0}(T) - Z_{t_0, x_0}(s)).$$

Therefore

$$u(s, x) = E \left(f(X_{t_0, x_0}(T)) \frac{Y_{t_0, x_0}(T)}{Y_{t_0, x_0}(s)} + \frac{1}{Y_{t_0, x_0}(s)}(Z_{t_0, x_0}(T) - Z_{t_0, x_0}(s)) \mid X_{t_0, x_0}(s) = x \right).$$

Thus, storing ${}_mX_{t_0, x_0}(t)$, ${}_mY_{t_0, x_0}(t)$, ${}_mZ_{t_0, x_0}(t)$, $t_0 \leq t \leq T$ (in fact, storing ${}_m\bar{X}$, ${}_m\bar{Y}$, ${}_m\bar{Z}$ at t_k), we get the pairs $({}_mX, {}_mV)$ from

$$(X, V) \sim \left(X_{t_0, x_0}(s), f(X_{t_0, x_0}(T)) \frac{Y_{t_0, x_0}(T)}{Y_{t_0, x_0}(s)} + \frac{1}{Y_{t_0, x_0}(s)}(Z_{t_0, x_0}(T) - Z_{t_0, x_0}(s)) \right).$$

Having this sample, one can obtain $\hat{u}(s, x)$ by the linear regression method (see section 2.4):

$$(3.8) \quad \hat{u}(s, x) = \sum_{l=1}^L \hat{\alpha}_l \varphi_l(x).$$

From (3.8) it is straightforward to obtain a very simple estimate $\hat{\partial}^i(s, x)$ for $\partial^i(s, x) = \partial u / \partial x^i(s, x)$:

$$(3.9) \quad \hat{\partial}^i(s, x) = \frac{\partial \hat{u}(s, x)}{\partial x^i} = \sum_{l=1}^L \hat{\alpha}_l \frac{\partial \varphi_l(x)}{\partial x^i}.$$

Then from (2.14) we find some $\hat{\mu}(s, x)$, $\hat{F}(s, x)$ for any $t_0 < s < T$ (in reality for any t_k) and construct the variate $\hat{\Gamma}(t_0, x_0)$ (see (2.5) and (2.7)) for $u(t_0, x_0)$ due to the system (2.10) with $\mu = \hat{\mu}$ and $F = \hat{F}$. We repeat that the variate $\hat{\Gamma}(t_0, x_0)$ is unbiased for any $\hat{\mu}$ and \hat{F} . We note that it is sufficient to have rather rough (in comparison with the required accuracy in evaluating $u(t_0, x_0)$) approximations $\hat{\mu}(s, x)$ and $\hat{F}(s, x)$ of some optimal μ and F from (2.13). Therefore, it is natural to use a coarser discretization and fewer MC runs in the regression part of evaluating $\hat{u}(s, x)$ due to (3.8), i.e., to take M_r in (2.22) smaller than M and to construct samples ${}_m X$ in (2.25) with a comparatively rough discretization. Then in computing $u(t_0, x_0)$ with a finer discretization, the necessary values of $\hat{\mu}$ and \hat{F} at the intermediate points can be obtained after, e.g., linear interpolation of \hat{u} with respect to time. The success of any regression-based approach clearly depends on the choice of basis functions. This is known to be a rather complicated problem, both in practice and theory. In fact, it is necessary to use a special basis tailored to each particular problem. Fortunately, the variance can easily be evaluated during simulation. Therefore, it is not very expensive from the computational point of view to check the quality of a given basis if we take coarse discretizations both in the regression part and in the main part of evaluating $u(t_0, x_0)$ and if we take not too large numbers M_r and M of MC runs. This can help in choosing a proper basis.

Remark 3.1. Clearly, $\hat{\alpha}_l$ depend on s (on t_k). Let us note that the number L and the set $\{\varphi_l(x)\}_{l=1}^L$ may depend on t_k as well.

Remark 3.2. It is obvious that in practice we use (2.10) with different μ and F in the implementation of the regression and in computing the required quantity $u(t_0, x_0)$. Indeed, in the regression part of the procedure we can take arbitrary μ and F (e.g., both zero), while in computing $u(t_0, x_0)$ we choose μ and F according to (2.14) with \hat{u} obtained via the regression or according to (2.15) with \hat{u} and $\widehat{\partial u} / \partial x^i$ obtained via the regression.

Remark 3.3. At $s = t_0$ the system (2.24) degenerates into the single equation (we suppose that not all of $\varphi_l(x_0)$ are equal to zero)

$$(3.10) \quad \varphi_1(x_0)\alpha_1 + \cdots + \varphi_L(x_0)\alpha_L = \frac{1}{M_r} \sum_{m=1}^{M_r} [f({}_m \bar{X}_{t_0, x_0}(T)) \quad {}_m \bar{Y}_{t_0, x_0}(T) + {}_m \bar{Z}_{t_0, x_0}(T)].$$

Therefore, the coefficients $\alpha_1(t_0), \dots, \alpha_L(t_0)$ cannot be found from (3.10) uniquely. At the same time, the linear combination $\alpha_1(t_0)\varphi_1(x_0) + \cdots + \alpha_L(t_0)\varphi_L(x_0)$, i.e., the estimate

$$\hat{u}(t_0, x_0) = \frac{1}{M_r} \sum_{m=1}^{M_r} [f({}_m \bar{X}_{t_0, x_0}(T)) \quad {}_m \bar{Y}_{t_0, x_0}(T) + {}_m \bar{Z}_{t_0, x_0}(T)],$$

is defined uniquely. Clearly, when t_k is close to t_0 (for instance, at t_1), the system (2.24), though not degenerate, is ill-conditioned. Nevertheless, for such t_k and for x

close to x_0 , the estimate

$$\hat{u}(t_k, x) = \alpha_1(t_k)\varphi_1(x) + \cdots + \alpha_L(t_k)\varphi_L(x)$$

can be found sufficiently accurate. However, since it is not possible to satisfactorily determine the coefficients $\alpha_1(t_k), \dots, \alpha_L(t_k)$, we cannot get the derivatives $\partial \hat{u}(t_k, x)/\partial x^i$ by direct differentiation as $\alpha_1(t_k)\partial \varphi_1(x)/\partial x^i + \cdots + \alpha_L(t_k)\partial \varphi_L(x)/\partial x^i$. In addition, let us emphasize that such difficulties are not essential for the whole procedure of variance reduction because the variance is equal to the integral (2.12), and unsatisfactory knowledge of u and $\partial u/\partial x^i$ on short parts of the interval $[t_0, T]$ does not significantly affect the value of the integral.

3.3. Evaluating $\partial u/\partial x^i(s, x)$. The problem of evaluating $\partial u/\partial x^i(s, x)$ is of independent importance due to its connection with numerical computation of Greeks in finance. Many articles are devoted to pathwise methods of estimating Greeks (see [7] and the references therein; see also [13]). In [17] the finite-difference-based method is developed, and [5, 4] suggest using Malliavin calculus for computing Greeks. Several pathwise and finite-difference-based methods for calculating sensitivities of Bermudan options using regression methods and MC simulations are considered in [1] (see also the references therein). In this section we propose a conditional version of the pathwise method, and in section 3.4 we present a conditional version of the approach based on the Malliavin integration by parts for evaluating $\partial u/\partial x^i(s, x)$.

As mentioned previously, differentiating the equality (3.8) gives an estimate for $\partial^i(s, x) = \partial u/\partial x^i(s, x)$ (see (3.9)); however, in general, it is rather rough. A more accurate way is to use the linear regression method directly.

In evaluating $\partial^i(s, x)$ by regression, the pair (X, V^i) has the form (see (3.2))

(3.11)

$$\begin{aligned} X &= X_{t_0, x_0}(s), \\ V^i &= \sum_{j=1}^d \frac{\partial f(X_{s,X}(T))}{\partial x^j} \delta_{s,X}^i X^j(T) Y_{s,X}(T) + f(X_{s,X}(T)) \delta_{s,X}^i Y(T) + \delta_{s,X}^i Z(T). \end{aligned}$$

We already have expressions for $X_{s,X}(T)$, $Y_{s,X}(T)$, $Z_{s,X}(T)$ via $X_{t_0, x_0}(t)$, $Y_{t_0, x_0}(t)$, $Z_{t_0, x_0}(t)$, with t being equal to s and T (see the formulas (3.4), (3.6), (3.7)). Our nearest aim is to express $\delta_{s,X}^i X^j(T)$, $\delta_{s,X}^i Y(T)$, $\delta_{s,X}^i Z(T)$ via $X_{t_0, x_0}(t)$, $Y_{t_0, x_0}(t)$, $Z_{t_0, x_0}(t)$, $\delta_{t_0, x_0}^i X^j(t)$, $\delta_{t_0, x_0}^i Y(t)$, $\delta_{t_0, x_0}^i Z(t)$.

We begin with $\delta_{s,X}^i X^j(t)$. The column-vector $\delta_{s,X}^i X(t)$ is the solution of the linear homogeneous stochastic system (2.17) whose coefficients depend on $X_{s,X}(t) = X_{t_0, x_0}(t)$. Let the matrix

$$\Phi_{s,X}(t) := \{\delta_{s,X}^i X^j(t)\}$$

be the fundamental matrix of solutions of (2.17) normalized at time s , i.e., $\Phi_{s,X}(s) = I$, where I is the identity matrix. Its element on the j th row and i th column is equal to $\delta_{s,X}^i X^j(t)$. Clearly,

$$(3.12) \quad \Phi_{s,X}(t) = \Phi_{t_0, x_0}(t) \Phi_{t_0, x_0}^{-1}(s).$$

Now let us turn to the column-vector $\delta_{s,X} Y(t)$, consisting of components $\delta_{s,X}^i Y(t)$. We have (see (2.18))

$$(3.13) \quad \begin{aligned} d\delta_{s,X} Y &= Y_{s,X}(t) \Phi_{s,X}^\top(t) \nabla c(t, X_{s,X}(t)) dt + c(t, X_{s,X}(t)) \delta_{s,X} Y dt \\ &+ Y_{s,X}(t) \Phi_{s,X}^\top(t) \nabla [\mu^\top(t, X_{s,X}(t)) dw(t)] + \delta_{s,X} Y \mu^\top(t, X_{s,X}(t)) dw(t), \quad \delta_{s,X} Y(s) = 0. \end{aligned}$$

Due to the equality $X_{s,X}(t) = X_{t_0,x_0}(t)$ and (3.6) and (3.12), we get from (3.13)

(3.14)

$$\begin{aligned} d\delta_{s,X}Y &= \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)} [\Phi_{t_0,x_0}^{-1}(s)]^\top \Phi_{t_0,x_0}^\top(t) \nabla c(t, X_{t_0,x_0}(t)) dt + c(t, X_{t_0,x_0}(t)) \delta_{s,X}Y dt \\ &\quad + \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)} [\Phi_{t_0,x_0}^{-1}(s)]^\top \Phi_{t_0,x_0}^\top(t) \nabla [\mu^\top(t, X_{t_0,x_0}(t)) dw(t)] \\ &\quad + \delta_{s,X}Y \mu^\top(t, X_{t_0,x_0}(t)) dw(t), \quad \delta_{s,X}Y(s) = 0. \end{aligned}$$

Taking into account the equality

$$\begin{aligned} d\delta_{t_0,x_0}Y(t) &= Y_{t_0,x_0}(t) \Phi_{t_0,x_0}^\top(t) \nabla c(t, X_{t_0,x_0}(t)) dt + c(t, X_{t_0,x_0}(t)) \delta_{t_0,x_0}Y(t) dt \\ &\quad + Y_{t_0,x_0}(t) \Phi_{t_0,x_0}^\top(t) \nabla [\mu^\top(t, X_{t_0,x_0}(t)) dw(t)] + \delta_{t_0,x_0}Y(t) \mu^\top(t, X_{t_0,x_0}(t)) dw(t), \end{aligned}$$

it is not difficult to verify that

$$(3.15) \quad \delta_{s,X}Y(t) = \frac{1}{Y_{t_0,x_0}(s)} [\Phi_{t_0,x_0}^{-1}(s)]^\top \left(\delta_{t_0,x_0}Y(t) - \frac{Y_{t_0,x_0}(t)}{Y_{t_0,x_0}(s)} \delta_{t_0,x_0}Y(s) \right).$$

In the similar way we obtain

$$\begin{aligned} (3.16) \quad \delta_{s,X}Z(t) &= \frac{1}{Y_{t_0,x_0}(s)} [\Phi_{t_0,x_0}^{-1}(s)]^\top (\delta_{t_0,x_0}Z(t) - \delta_{t_0,x_0}Z(s)) \\ &\quad - \frac{1}{Y_{t_0,x_0}^2(s)} [\Phi_{t_0,x_0}^{-1}(s)]^\top \delta_{t_0,x_0}Y(s) (Z_{t_0,x_0}(t) - Z_{t_0,x_0}(s)). \end{aligned}$$

Hence the column-vector $\partial(s, x)$ with the components $\partial^i(s, x)$ is equal to

$$\begin{aligned} (3.17) \quad \partial(s, x) &= E \left(\frac{Y_{t_0,x_0}(T)}{Y_{t_0,x_0}(s)} [\Phi_{t_0,x_0}^{-1}(s)]^\top \Phi_{t_0,x_0}^\top(T) \nabla f(X_{t_0,x_0}(T)) \right. \\ &\quad \left. + f(X_{t_0,x_0}(T)) \delta_{s,X}Y(T) + \delta_{s,X}Z(T) \mid X_{t_0,x_0}(s) = x \right), \end{aligned}$$

where $\delta_{s,X}Y(T)$ and $\delta_{s,X}Z(T)$ are from (3.15) and (3.16).

Thus, storing ${}_mX_{t_0,x_0}(t)$, ${}_mY_{t_0,x_0}(t)$, ${}_mZ_{t_0,x_0}(t)$, ${}_m\Phi_{t_0,x_0}(t)$, ${}_m\delta_{t_0,x_0}Y(t)$, ${}_m\delta_{t_0,x_0}Z(t)$, $t_0 \leq t \leq T$, we get the corresponding samples

(3.18)

$$\begin{aligned} ({}_mX, {}_mV^i) &= \left({}_mX_{t_0,x_0}(s), \left(\frac{{}_mY_{t_0,x_0}(T)}{{}_mY_{t_0,x_0}(s)} [{}_m\Phi_{t_0,x_0}^{-1}(s)]^\top {}_m\Phi_{t_0,x_0}^\top(T) \nabla f({}_mX_{t_0,x_0}(T)) \right. \right. \\ &\quad \left. \left. + f({}_mX_{t_0,x_0}(T)) {}_m\delta_{s,{}_mX}Y(T) + {}_m\delta_{s,{}_mX}Z(T) \right)^i \right), \end{aligned}$$

where ${}_m\Phi_{t_0,x_0}(s)$ is a realization of the fundamental matrix $\Phi_{t_0,x_0}(s)$ which corresponds to the same elementary event $\omega \in \Omega$ as the realization ${}_mX_{t_0,x_0}(t)$. We use $({}_mX, {}_mV^i)$ for evaluating $\partial^i(s, x)$, $i = 1, \dots, d$, by the linear regression method:

$$(3.19) \quad \hat{\partial}^i(s, x) = \sum_{l=1}^L \hat{\beta}_l^i \psi_l(x).$$

Remark 3.4. This paper is most closely connected with [6, 12, 13, 14] (see also [16]) and with the works [21, 20] by N. Newton. In [21, 20], both the method of control variates and the method of importance sampling for calculating solutions $u(t, x)$ of parabolic partial differential equations by the MC method are considered. In both cases, a perfect variate (i.e., one which is unbiased and has zero variance) is constructed based on the Funke–Shevlyakov–Haussmann formula (see the corresponding reference and details in [21]; such a formula is usually called as the Clark–Ocone–Haussmann formula). Then some approximation methods of simulating the variates are proposed in [21, 20] to yield unbiased estimators for the desired solution $u(t, x)$ with reduced variances. If the dimension d is large, the most labor-consuming calculations are connected with integration of the d^2 -dimensional system of first-order variation equations. This is required to construct the estimators. In this paper, we use variates in the form (2.11), (2.10) with μ and F satisfying (2.13). Due to Theorem 2.1, these variates are perfect if u and $\partial u/\partial x^i$ are exact. We evaluate u and $\partial u/\partial x^i$ based on conditional probabilistic representations and construct unbiased estimators for $u(t, x)$ using (2.15) or (2.14). We note that (2.14) allows us to avoid estimating $\partial u/\partial x^i$ (see (3.8)–(3.9)) and hence to avoid integration of the equations of first-order variation. In addition, the obtained estimator by (2.14) remains unbiased. In spite of the fact that our approach and that of N. Newton clearly differ, they undoubtedly have profound connections. For example, the Clark–Ocone–Haussmann formula, being the basis for Newton’s approach, can fairly easily be derived using the conditional probabilistic representations (3.1), (3.2).

3.4. Evaluating $\partial u/\partial x^i(s, x)$ using the Malliavin integration by parts.

If $f(x)$ is an irregular function, one can use the procedure recommended in section 3.2, where we do not need direct calculations of derivatives $\partial u/\partial x^i$. Another way consists in approximating f by a smooth function with the consequent use of the procedure from section 3.3. Because we do not pursue a high accuracy in estimating u and $\partial u/\partial x^i$, such approximation of f can be quite satisfactory. For direct calculation of derivatives $\partial u/\partial x^i$ without smoothing f , we can use the conditional version of the integration-by-parts (Bismut–Elworthy–Li) formula. This formula is successfully applied for evaluating deltas in the case of an irregular f (see, e.g., [5, 4, 22]).

For calculating $\partial u/\partial x^i$ in the case of u given by

$$u(s, x) = E\Gamma_{s,x} = E[f(X_{s,x}(T))Y_{s,x}(T) + Z_{s,x}(T)],$$

where $X_{s,x}(T)$, $Y_{s,x}(T)$, $Z_{s,x}(T)$ satisfy system (2.10), the following variant of the integration-by-parts formula can be derived:

$$\begin{aligned} (3.20) \quad \partial^i(s, x) &= \frac{1}{T-s} E\Gamma_{s,x} \int_s^T \left[\sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} \right]^\top dw(s') \\ &\quad - \frac{1}{T-s} E\Gamma_{s,x} \int_s^T \mu^\top \sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} ds' + \frac{1}{T-s} E \int_s^T Z_{s,x}(s') \mu^\top \sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} ds' \\ &\quad + \frac{1}{T-s} E\Gamma_{s,x} \int_s^T \frac{1}{Y_{s,x}(s')} \frac{\partial Y_{s,x}(s')}{\partial x^i} ds' - \frac{1}{T-s} E \int_s^T \frac{Z_{s,x}(s')}{Y_{s,x}(s')} \frac{\partial Y_{s,x}(s')}{\partial x^i} ds' \\ &\quad - \frac{1}{T-s} E \int_s^T Y_{s,x}(s') F^\top \sigma^{-1} \frac{\partial X_{s,x}(s')}{\partial x^i} ds' + \frac{1}{T-s} E \int_s^T \frac{\partial Z_{s,x}(s')}{\partial x^i} ds' := D^i(s, x), \end{aligned}$$

where μ^\top , σ^{-1} , and F^\top have $(s', X_{s,x}(s'))$ as their arguments. In particular, if $c = 0$, $g = 0$, $\mu = 0$, $F = 0$, we get the well-known integration-by-parts formula (see,

e.g., [22]):

$$(3.21) \quad \partial^i(s, x) = \frac{1}{T-s} Ef(X_{s,x}(T)) \int_s^T \left[\sigma^{-1}(s', X_{s,x}(s')) \frac{\partial X_{s,x}(s')}{\partial x^i} \right]^\top dw(s').$$

As in section 3.1, together with the unconditional probabilistic representation (3.20) for $\partial^i(s, x)$, we have the following conditional one:

$$(3.22) \quad \partial^i(s, x) = E(D^i(s, X)|X := X_{t_0, x_0}(s) = x).$$

Again, the formula (3.20) is natural for the MC approach and (3.22) for a regression method. An implementation of the regression method is based upon the corresponding approximation $({}_mX, {}_mV^i)$ of the pair $(X, V^i) = (X_{t_0, x_0}(s), D^i(s, X_{t_0, x_0}(s)))$ following the ideas of section 3.3.

3.5. Two-run procedure. The straightforward implementation of evaluating $u(s, x)$ and $\partial u/\partial x^i(s, x)$ by regression as described in sections 3.2 and 3.3 requires storing

$${}_m\Lambda(t_k) := ({}_mX_{t_0, x_0}(t_k), {}_mY_{t_0, x_0}(t_k), {}_mZ_{t_0, x_0}(t_k), {}_m\Phi_{t_0, x_0}(t_k), {}_m\delta_{t_0, x_0}Y(t_k), {}_m\delta_{t_0, x_0}Z(t_k))$$

(or, more precisely, their approximations ${}_m\bar{\Lambda}(t_k)$) at all t_k , $k = 1, \dots, N$, in the main computer memory (RAM) until the end of the simulation. This puts a requirement on the RAM size that is too demanding and limits the practicality of the proposed approach since in almost any practical problem a relatively large number of time steps is needed. However, this difficulty can be overcome and we can avoid storing ${}_m\bar{\Lambda}(t_k)$ at all t_k by implementing the two-run procedure described below.

First, we recall that, as a rule, pseudorandom number generators used for MC simulations have the property that the sequence of random numbers obtained by them is easily reproducible (see, e.g., [16] and the references therein). Let us fix a sequence of pseudorandom numbers. The two-run procedure can schematically be presented as follows.

First run:

- simulate M_r number of independent trajectories ${}_m\bar{\Lambda}(t_k)$, $k = 1, \dots, N$, with an arbitrary choice of μ and F (e.g., $\mu = 0$ and $F = 0$);
- compute and store the values ${}_m\bar{\Gamma}$ to form the component V needed for the regression in the second run and compute and store the values

$${}_m\bar{Y}(T) {}_m\bar{\Phi}_{t_0, x_0}^\top(T) \nabla f({}_m\bar{X}(T)) + f({}_m\bar{X}(T)) {}_m\bar{\delta Y}(T) + {}_m\bar{\delta Z}(T)$$

and ${}_m\bar{Y}(T)$ to form the components V^i in the second run.

Second run:

- reinitialize the random number generator so that it produces the same sequence as for the first run;
- for $k = 1, \dots, N$
 - simulate the same ${}_m\bar{\Lambda}(t_k)$, $m = 1, \dots, M_r$, as in the first run (i.e., they correspond to the same sequence of pseudorandom numbers as in the first run), keeping only the current ${}_m\bar{\Lambda}(t_k)$ in RAM;
 - use the values stored in RAM during the first run and ${}_m\bar{\Lambda}(t_k)$ from this run to find $\bar{u}(t_k, x)$ and $\bar{\partial u}/\partial x^i(t_k, x)$ by regression (${}_m\bar{\Lambda}(t_k)$ and ${}_m\bar{\Lambda}(T)$ form the pairs $({}_mX, {}_mV)$ and $({}_mX, {}_mV^i)$ needed for the regression);

- use the found $\bar{u}(t_k, x)$ and $\overline{\partial u / \partial x^i}(t_k, x)$ to obtain $\bar{\mu}(t_k, x)$ and $\bar{F}(t_k, x)$ required for variance reduction (see section 2.2);
- simulate (2.10) with $\mu = \bar{\mu}$ and $F = \bar{F}$ on this step and thus obtain M independent triples

$$({}_m\tilde{X}_{t_0, x_0}(t_k), {}_m\tilde{Y}_{t_0, x_0}(t_k), {}_m\tilde{Z}_{t_0, x_0}(t_k)) = ({}_m\tilde{X}_{t_{k-1}, m\tilde{X}(t_{k-1})}(t_k), {}_m\tilde{Y}_{t_{k-1}, m\tilde{X}(t_{k-1})}(t_k), {}_m\tilde{Z}_{t_{k-1}, m\tilde{X}(t_{k-1})}(t_k)),$$

which we keep in RAM until the next step;

- use the obtained $({}_m\tilde{X}_{t_0, x_0}(T), {}_m\tilde{Y}_{t_0, x_0}(T), {}_m\tilde{Z}_{t_0, x_0}(T))$ to get the required $u(t_0, x_0)$ (see (2.6)).

We emphasize that in the two-run procedure at each time moment $s = t_k$ we need to keep in memory only the precomputed values stored at the end of the first run and the values ${}_m\tilde{X}(t_k)$ and $({}_m\tilde{X}_{t_0, x_0}(t_k), {}_m\tilde{Y}_{t_0, x_0}(t_k), {}_m\tilde{Z}_{t_0, x_0}(t_k))$ (only at the current time step k), which is well within RAM limits of a PC.

We note that the two-run realization of the procedure from section 3.2 based on using regression for estimating u only is less computationally demanding (both on processor time and RAM and especially for problems of large dimension d) than the procedures of sections 3.3 and 3.4 which estimate the derivatives of u via regression.

The two-run procedure was used in the numerical experiments of sections 4.2 and 4.3.

4. Examples. The first example is partly illustrative and partly theoretical. The second and third examples are numerical.

4.1. Heat equation. Consider the Cauchy problem

$$(4.1) \quad \frac{\partial u}{\partial t} + \frac{\sigma^2}{2} \frac{\partial^2 u}{\partial x^2} = 0, \quad t_0 \leq t < T, \quad x \in \mathbf{R},$$

$$u(T, x) = x^2.$$

Its solution is

$$(4.2) \quad u(t, x) = \sigma^2(T - t) + x^2.$$

The probabilistic representation (2.10), (2.11) with $\mu = 0$ takes the form

$$(4.3) \quad u(s, x) = E[X_{s,x}^2(T) + Z_{s,x}(T)] = E\Gamma_{s,x},$$

$$(4.4) \quad dX = \sigma dw(t), \quad X(s) = x,$$

$$(4.5) \quad dZ = F(t, X)dw(t), \quad Z(s) = 0.$$

Due to Theorem 2.1, we have $\text{var}\Gamma_{s,x} = \text{var}[X_{s,x}^2(T) + Z_{s,x}(T)] = 0$ for the optimal choice of the function $F(t, x) = -\sigma \partial u / \partial x = -2\sigma x$. We note that in this example $\partial u / \partial x$ and the optimal F do not depend on time t .

For the purpose of this illustrative example, we evaluate $u(0, 0) = E\Gamma_{0,0}$. Let us simulate (4.4) exactly (i.e., we have no error of numerical integration):

$$(4.6) \quad X_0 = x, \quad X_{k+1} = X_k + \sigma \Delta_k w, \quad k = 0, \dots, N-1, \quad \Delta_k w := w(t_{k+1}) - w(t_k).$$

For $F \equiv 0$, we have $u(0, 0) = E\Gamma_{0,0} \approx \hat{u}(0, 0) = \frac{1}{M} \sum_{m=1}^M {}_mX_N^2$, where ${}_mX_N$ are independent realizations of X_N obtained by (4.6). Further, $\text{var}\Gamma_{0,0} = 2\sigma^4 T^2$, and

hence the MC error is equal to (see (2.9))

$$(4.7) \quad \rho = c \frac{\sqrt{2}\sigma^2 T}{\sqrt{M}}.$$

For instance, to achieve the accuracy $\rho = 0.0001$ for $c = 3$ (recall that there is no error of numerical integration here) in the case of $\sigma = 1$ and $T = 10$, one needs to perform $M = 18 \times 10^{10}$ MC runs.

To reduce the MC error, we estimate $\partial u / \partial x$ by regression to get $\hat{F}(t_k, x)$ close to the optimal $F = -2\sigma x$. As the basis functions for the regression, we take the first two Hermite polynomials:

$$(4.8) \quad \psi_1(x) = 1, \quad \psi_2(x) = 2x.$$

We note that in this example the required derivative $\partial u / \partial x$ can be expanded in the basis (4.8); i.e., here we do not have any error due to the cut-off of a set of basis functions. In the construction of the estimate for $\partial u / \partial x$, we put $F = 0$ in (4.5).

The variational equation associated with (4.4) has the form (see (2.17)) $d\delta X = 0$, $\delta X(s) = 1$, and hence $\delta X(t) = 1$, $t \geq s$. Thus, the sample from (3.18) takes the form $({}_m X, {}_m V) = ({}_m X_{t_0, x_0}(s), 2 {}_m X_{t_0, x_0}(T))$ and the estimator $\hat{\partial}(t_k, x)$ for $\partial u / \partial x(t_k, x)$ is constructed as

$$(4.9) \quad \hat{\partial}(t_k, x) = \hat{\alpha}_1(t_k) + 2\hat{\alpha}_2(t_k)x, \quad k = 1, \dots, N,$$

where $\hat{\alpha}_1(t_k)$ and $\hat{\alpha}_2(t_k)$ satisfy the system of linear algebraic equations (see (2.24)–(2.25))

$$(4.10) \quad \begin{aligned} a_{11}\alpha_1 + a_{12}\alpha_2 &= b_1, \\ a_{21}\alpha_1 + a_{22}\alpha_2 &= b_2, \end{aligned}$$

$$(4.11) \quad a_{11} = 1, \quad a_{12} = a_{21} := a_{12}(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 2 \times {}_m X(t_k),$$

$$a_{22} := a_{22}(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 4 \times ({}_m X(t_k))^2,$$

$$b_1 := b_1(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 2 \times {}_m X(T), \quad b_2 := b_2(t_k) = \frac{1}{M_r} \sum_{m=1}^{M_r} 4 \times {}_m X(t_k) \times {}_m X(T).$$

Here ${}_m X(t_k)$, $m = 1, \dots, M_r$, $k = 1, \dots, N$, are independent realizations of $X(t_k)$ obtained by (4.6). Hence

$$(4.12) \quad \hat{\alpha}_1(t_k) = \frac{b_1 a_{22} - b_2 a_{12}}{a_{22} - (a_{12})^2}, \quad \hat{\alpha}_2(t_k) = \frac{b_2 - b_1 a_{12}}{a_{22} - (a_{12})^2}.$$

We define

$$(4.13) \quad \hat{F}(0, x) = -\frac{\sigma}{M_r} \sum_{m=1}^{M_r} 2 \times {}_m X(T),$$

$$\hat{F}(t, x) = -\sigma (\hat{\alpha}_1(t_k) + 2\hat{\alpha}_2(t_k)x) \quad \text{for } t \in (t_{k-1}, t_k], \quad k = 1, \dots, N.$$

We simulate (4.5) with $F = \hat{F}(t, x)$ exactly (i.e., again we have no error of numerical integration):

$$(4.14) \quad Z_0 = 0,$$

$$Z_{k+1} = Z_k - \sigma \hat{\alpha}_1(t_{k+1}) \Delta_k w - 2\sigma^2 \hat{\alpha}_2(t_{k+1}) w(t_k) \Delta_k w - \sigma^2 \hat{\alpha}_2(t_{k+1}) \left[(\Delta_k w)^2 - h \right].$$

The increments $\Delta_k w$ are the same both in (4.6) and in (4.14) and are independent of the ones used to estimate $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

We simulate

$$(4.15) \quad u(0, 0) = E\Gamma_{0,0} = E(X_N^2 + Z_N) \approx \hat{u}(0, 0) = \frac{1}{M_r} \sum_{m=1}^{M_r} ({}_m X_N^2 + {}_m Z_N),$$

where ${}_m X_N$ and ${}_m Z_N$ are independent realizations of X_N and Z_N obtained according to (4.6) and (4.14). We note that the approximation (4.15) does not have the numerical integration error or the error due to the cut-off of the basis; it has the MC error only.

Using Theorem 2.1, one can evaluate $\text{var}\Gamma_{0,0}$ in the case of $F = \hat{F}$ defined in (4.13) and obtain $\text{var}\Gamma_{0,0} \approx 4\sigma^4 T^2 / M_r$. Then the MC error ρ in this case is equal to (compare with (4.7))

$$(4.16) \quad \rho \approx c \frac{2\sigma^2 T}{\sqrt{M M_r}}.$$

This example illustrates that in the absence of the error due to the cut-off of a set of basis functions used in regression and of the numerical integration error, the MC error is reduced $\sim 1/\sqrt{M_r}$ times by the proposed variance reduction technique. This is, of course, a significant improvement. Indeed, let us return to the example discussed after (4.7). The estimate (4.16) implies that to achieve the accuracy $\rho = 0.0001$ for $c = 3$ in the case of $\sigma = 1$ and $T = 10$, one can take, e.g., $M = M_r = 6 \times 10^5$; i.e., one can run about 10^5 times fewer trajectories than when the variance reduction was not used (see the discussion after (4.7)). The gain of computational efficiency is significant in spite of the fact that there is an overhead cost of solving the linear system (4.10) in the “regression’s runs.”

Remark 4.1. In the above analysis we assumed that “regression’s runs” and the MC runs for computing the desired value $u(0, 0)$ are independent. In practice, this assumption can be dropped, and we can use the same paths $X(t)$ for both the “regression’s runs” and the MC runs. Then, as a rule, we choose $M_r \leq M$.

Remark 4.2. We are expecting (see also experiments in section 4.2) that in the general case the MC error after application of this variance reduction technique has the form

$$(4.17) \quad \rho = O\left(\frac{1}{\sqrt{M M_r}} + \frac{h^{p/2}}{\sqrt{M}} + \frac{\text{err}_B}{\sqrt{M}}\right),$$

where the first term has the same nature as in this illustrative example (see (4.16)); the second term is due to the error of numerical integration (it is assumed that a method of weak order p is used); and the third one arises as a result of the use of a finite set of functions as the basis in the regression, while the solution $u(t, x)$ is usually expandable in a basis consisting of an infinite number of functions (i.e., this

error is due to the cut-off of the basis). We note that finding an appropriate basis for regression in applying this variance reduction approach to a particular problem can be a difficult task and requires some knowledge of the solution $u(t, x)$ of the considered problem. Roughly speaking, in the proposed implementation of the variance reduction methods (the method of importance sampling, the method of control variates, or the combining method) we substitute the task of finding an approximate solution to the problem of interest with the task of finding an appropriate basis for the regression.

For complicated systems of SDEs, it is preferable to use regression to approximate the solution $u(t, x)$ and then differentiate this approximation to approximate the derivatives $\partial u / \partial x^i$. In the case of this illustrative example we take the first three Hermite polynomials,

$$(4.18) \quad \psi_1(x) = 1, \quad \psi_2(x) = 2x, \quad \psi_3(x) = 4x^2 - 2,$$

as the basis functions for the regression. In this example the required function $u(t, x)$ can be expanded in the basis (4.18). We construct the estimator $\hat{u}(t_k, x)$ for $u(t_k, x)$:

$$(4.19) \quad \hat{u}(t_k, x) = \hat{\alpha}_1(t_k) + 2\hat{\alpha}_2(t_k)x + \hat{\alpha}_3(t_k) \cdot (4x^2 - 2), \quad k = 1, \dots, N,$$

where $\hat{\alpha}_1(t_k)$, $\hat{\alpha}_2(t_k)$, $\hat{\alpha}_3(t_k)$ satisfy the system of linear algebraic equations (2.24) with the corresponding coefficients. Further, we approximate the derivative $\partial u / \partial x(t_k, x)$,

$$(4.20) \quad \frac{\partial u}{\partial x}(t_k, x) \approx 2\hat{\alpha}_2(t_k) + 8\hat{\alpha}_3(t_k)x,$$

with $\hat{\alpha}_2(t_k)$ and $\hat{\alpha}_3(t_k)$ from (4.19), and we define

$$(4.21) \quad \hat{F}(t, x) := -\sigma(2\hat{\alpha}_2(t_k) + 8\hat{\alpha}_3(t_k)x) \quad \text{for } t \in [t_{k-1}, t_k], \quad k = 1, \dots, N,$$

which we use for variance reduction by putting $F = \hat{F}$ in (4.5). In the experiments we simulate (4.5) with $F = \hat{F}(t, x)$ exactly (see (4.14)). The new estimator for $u(0, 0)$ has the form (4.15) again but with the new Z_N corresponding to the choice of $\hat{F}(t, x)$ from (4.21).

TABLE 1

Heat equation. Simulation of $u(0, 0)$ for $\sigma = 1$ and $T = 10$ by (4.15) with the corresponding choice of the function F and for various M . The time step $h = 0.1$ and $M_r = M$. The exact value is $u(0, 0) = 10$. The value after “ \pm ” equals two standard deviations of the corresponding estimator and gives the confidence interval for the corresponding value with probability 0.95 (i.e., $c = 2$).

M	$F = 0$	$F = \hat{F}$ from (4.13)	$F = \hat{F}$ from (4.21)
10^3	9.67 ± 0.85	9.993 ± 0.045	9.999 ± 0.101
10^4	9.92 ± 0.28	9.9970 ± 0.0058	9.999 ± 0.012
10^5	9.970 ± 0.089	10.0000 ± 0.0003	10.0014 ± 0.0014

Table 1 gives some results of simulating $u(0, 0)$ by (4.15) with $F = 0$, $F = \hat{F}$ from (4.13), and $F = \hat{F}$ from (4.21). We see that for $F = 0$ the MC error is consistent with (4.7); i.e., it decreases $\sim 1/\sqrt{M}$. When the variance reduction is used, the results in Table 1 approve the MC error estimate (4.16). It is quite obvious that \hat{F} from (4.13) is a more accurate estimator for the exact $F = -2\sigma x$ than \hat{F} from (4.21), and then the MC error in the first case should usually be less than in the second case, which is observed in the experiments as well.

We also did similar experiments in the case of the terminal condition $u(T, x) = x^4$ in (4.1). To estimate $\partial u / \partial x$ by regression, we took the basis consisting of the first four Hermite polynomials. The results were analogous to those given above for the case x^2 .

4.2. Ergodic limit for one-dimensional array of stochastic oscillators.

Consider the one-dimensional array of oscillators [23, 19]:

(4.22)

$$\begin{aligned} dP^i &= -V'(Q^i) dt - \lambda \cdot (2Q^i - Q^{i+1} - Q^{i-1}) dt - \nu P^i dt + \sigma dw_i(t), \quad P^i(0) = p^i, \\ dQ^i &= P^i dt, \quad Q^i(0) = q^i, \quad i = 1, \dots, n, \end{aligned}$$

where periodic boundary conditions are assumed, i.e., $Q^0 := Q^n$ and $Q^{n+1} := Q^1$; $w_i(t)$, $i = 1, \dots, n$, are independent standard Wiener processes; $\nu > 0$ is a dissipation parameter; $\lambda \geq 0$ is a coupling constant; σ is the noise intensity; and $V(z)$, $z \in \mathbf{R}$, is a potential.

The SDEs (4.22) are ergodic with the Gibbs invariant measure μ . We are interested in computing the average of the potential energy with respect to the invariant measure associated with (4.22):

$$E_\mu U(Q) = E_\mu \sum_{i=1}^n \left(V(Q^i) + \frac{\lambda}{2} \cdot (Q^i - Q^{i+1})^2 \right).$$

To this end (see further details in [19]), we simulate the system (4.22) on a long time interval and approximate the ergodic limit $E_\mu U(Q)$ by $EU(Q(T))$ for a large T . To illustrate variance reduction via regression, we simulate

$$(4.23) \quad u(0, p, q) = EU(Q_{p,q}(T)) = E[U(Q_{p,q}(T)) + Z_{p,q}(T)],$$

where $Z(t)$, $0 \leq t \leq T$, satisfies

$$(4.24) \quad dZ = F^\top(t, P, Q)dw(t), \quad Z(0) = 0.$$

We choose the n -dimensional vector function $F(t, p, q)$ to be equal to (see (2.14))

$$(4.25) \quad F^i(t, p, q) = -\sigma \frac{\partial \hat{u}}{\partial p^i}, \quad i = 1, \dots, n,$$

where $\hat{u} = \hat{u}(t, p, q)$ is an approximation of the function

$$u(t, p, q) := EU(Q_{t,p,q}(T)).$$

We simulate (4.22) using the second-order weak quasi-symplectic integrator from [15, 16]:

$$(4.26) \quad P_0 = p, \quad Q_0 = q,$$

$$\begin{aligned} \mathcal{P}_{1,k}^i &= e^{-\nu h/2} P_k^i, \quad \mathcal{Q}_{1,k}^i = Q_k^i + \frac{h}{2} \mathcal{P}_{1,k}^i, \\ \mathcal{P}_{2,k}^i &= \mathcal{P}_{1,k}^i + h \left\{ -V'(\mathcal{Q}_{1,k}^i) - \lambda \cdot (2\mathcal{Q}_{1,k}^i - \mathcal{Q}_{1,k}^{i+1} - \mathcal{Q}_{1,k}^{i-1}) \right\} + h^{1/2} \sigma \xi_{ik}, \\ \mathcal{P}_{k+1}^i &= e^{-\nu h/2} \mathcal{P}_{2,k}^i, \quad \mathcal{Q}_{k+1}^i = \mathcal{Q}_{1,k}^i + \frac{h}{2} \mathcal{P}_{2,k}^i, \quad i = 1, \dots, n, \quad k = 0, \dots, N-1, \end{aligned}$$

where ξ_{ik} are independent and identically distributed random variables with the law

$$(4.27) \quad P(\xi = 0) = 2/3, \quad P(\xi = \pm\sqrt{3}) = 1/6.$$

And we approximate (4.24) by the standard second-order weak method (see [16, p. 103]):

(4.28)

$$\begin{aligned}
 Z_0 &= 0, \\
 Z_{k+1} &= Z_k + h^{1/2} \sum_{i=1}^n F^i(t_k, P_k, Q_k) \xi_{ik} + \sigma h \sum_{r=1}^n \sum_{i=1}^n \frac{\partial}{\partial p^i} F^r(t_k, P_k, Q_k) \xi_{irk} \\
 &\quad + \frac{1}{2} h^{3/2} \sum_{i=1}^n \mathcal{L} F^i(t_k, P_k, Q_k) \xi_{ik}, \\
 \xi_{irk} &= \frac{1}{2} \xi_{ik} \xi_{rk} - \frac{1}{2} \gamma_{ir} \zeta_{ik} \zeta_{rk}, \quad \gamma_{ir} = \begin{cases} -1, & i < r, \\ 1, & i \geq r, \end{cases} \\
 \mathcal{L} &:= \frac{\partial}{\partial t} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2}{\partial p^i \partial p^j} + \sum_{i=1}^n (-V'(q^i) - \lambda \cdot (2q^i - q^{i+1} - q^{i-1}) - \nu p^i) \frac{\partial}{\partial p^i} \\
 &\quad + \sum_{i=1}^n p^i \frac{\partial}{\partial q^i},
 \end{aligned}$$

where ξ_{ik} and ζ_{jk} are mutually independent random variables, ξ_{ik} are distributed by the law (4.27), and the ζ_{ik} are distributed by the law $P(\zeta = \pm 1) = 1/2$.

We consider two potentials: the harmonic potential

$$(4.29) \quad V(z) = \frac{1}{2} z^2, \quad z \in \mathbf{R},$$

and the hard anharmonic potential

$$(4.30) \quad V(z) = \frac{1}{2} z^2 + \frac{1}{2} z^4, \quad z \in \mathbf{R}.$$

We define the approximation $\hat{u}(t, p, q)$ used in (4.25) at $t = t_k$, $k = 0, \dots, N-1$, as follows. First, it is reasonable to put $\partial \hat{u} / \partial p^i(t, p, q) = 0$ for $0 \leq t \leq T_0$ with some relatively small T_0 since for large T the function $u(t, p, q)$, $0 \leq t \leq T_0$, is almost constant due to the ergodicity (the expectation in (4.23) is almost independent of the initial condition).

Further, let T_0 , T , h , N , and a nonnegative integer \varkappa be such that $T_0 = N_0 h$, $T = Nh$, $N - N_0 = \varkappa N'$, where N_0 and N' are integers. Introduce $\theta_{k'} = t_{N_0+k'\varkappa}$, $k' = 1, \dots, N'$.

In the case of harmonic potential the required function $u(t, p, q)$ can be expanded in the basis consisting of the finite number of functions

$$(4.31) \quad \varphi_l \in \{1, p^i, q^i, p^i p^j, q^i q^j, p^i q^j, \quad i, j = 1, \dots, n\}.$$

In our experiments we deal with three oscillators ($n = 3$); the basis (4.31) in this case has 28 functions.

We use the set of functions (4.31) as a set of basis functions for regression in both cases of harmonic and hard anharmonic potentials. Namely, using regression as described in section 3.2, we construct the estimator $\hat{u}(\theta_{k'}, p, q)$ for $u(\theta_{k'}, p, q)$ as

$$(4.32) \quad \hat{u}(\theta_{k'}, p, q) = \sum_{l=1}^L \hat{\alpha}_l(\theta_{k'}) \varphi_l(p, q),$$

where φ_l are defined in (4.31) and $\hat{\alpha}_l(\theta_{k'})$ satisfy the system of linear algebraic equations (2.24). The matrix formed from $\hat{\alpha}_l(\theta_{k'})$ is positive definite, and we solve the system of linear algebraic equations by Cholesky decomposition. To find the estimator \hat{u} , we use M_r independent trajectories.

Then for $T_0 < t_k < T$ we put $\hat{u}(t_k, p, q) = \hat{u}(\theta_{k'}, p, q)$ with $\theta_{k'} \leq t_k < \theta_{k'+1}$. The recalculation of the estimator \hat{u} once per a few number of steps \varkappa reduces the cost of the procedure.

We note that for the basis (4.31) the corresponding function F from (4.25) is such that some terms in the scheme (4.28) are canceled; in particular, it is not required to simulate the ζ_{ik} in this case.

We compute $u(0, p, q)$ in the usual way,

$$(4.33) \quad \begin{aligned} u(0, p, q) &= E[U(Q_{p,q}(T)) + Z_{p,q}(T)] \approx E[U(Q_N) + Z_N] \\ &\approx \frac{1}{M} \sum_{m=1}^M [U(mQ_N) + {}_mZ_N], \end{aligned}$$

by simulating M independent realizations of Q_N , Z_N from (4.26), (4.28). In these experiments the two-run procedure described in section 3.5 was used.

Suppose we would like to compute $u(0, p, q)$ for the particular set of parameters $n = 3, \lambda = 1, \nu = 1, \sigma = 1, T = 10$ and the potentials (4.29) and (4.30) with accuracy of order 10^{-3} . Since we are using the scheme of order two, we can take $h = 0.02$.

Let us first consider the case of harmonic potential (4.29). Without variance reduction (i.e., for $F = 0$), we obtain 0.7500 ± 0.0010 with the fiducial probability 95% by simulating $M = 1.4 \times 10^6$ trajectories, taking ~ 541 sec on a PC. When we use the variance reduction technique as described above, it is sufficient to take $T_0 = 2, \varkappa = 2, M_r = 2 \times 10^4, M = 3 \times 10^4$ to get 0.7496 ± 0.0010 in ~ 64 sec. In this example the procedure with variance reduction requires an eighth of the computational time. All the expenses are taken into account, including the time required for the first run of the two-run procedure, which is less than 10% of the total time. We recall that in this case the required function $u(t, p, q)$ can be expanded in the finite basis (4.31), unlike the case of hard anharmonic potential when such a basis is infinite.

Now consider the case of hard anharmonic potential (4.30). Without variance reduction (i.e., for $F = 0$), we obtain 0.6491 ± 0.0011 with the fiducial probability 95% by simulating $M = 10^6$ trajectories, taking ~ 403 sec on a PC. With variance reduction, we reach the same level of accuracy 0.6491 ± 0.0011 in ~ 98 sec by choosing, e.g., $T_0 = 2, \varkappa = 2, M_r = 2.5 \times 10^4, M = 5.5 \times 10^4$. Thus, the procedure with variance reduction requires a quarter of the computational time.

Some other results of our numerical experiments are presented in Tables 2 and 3. They show dependence of the MC error on M and M_r . The numerical integration error is relatively small here and does not essentially affect the results. The case $M_r = 0$ means that the simulation was done without variance reduction. We observe that in both tables for a fixed M_r the MC error decreases $\sim 1/\sqrt{M}$. Further, we see from Table 2 that the MC error is $\sim 1/\sqrt{M_r}$ for fixed M (for $M_r > 0$, of course), and, consequently, it is $\sim 1/\sqrt{MM_r}$ when the variance reduction is used (we recall that the time step is relatively small here). As noted before, the basis used in the variance reduction is such that the function $u(t, x)$ can be expanded in it in the case of harmonic potential; i.e., err_B in (4.17) is equal to 0. These observations are consistent with the MC error estimate (4.17). For the anharmonic potential, err_B is not equal to zero, and we see in Table 3 that the increase of M_r has less impact on the MC error in this case.

TABLE 2

Harmonic potential. Two standard deviations of the estimator (4.33) in the case of potential (4.29) for different M and M_r . $M_r = 0$ means that variance reduction was not used. The other parameters are $n = 3$, $\lambda = 1$, $\nu = 1$, $\sigma = 1$, $T = 10$ and $h = 0.01$, $T_0 = 2$, $\varkappa = 1$.

	$M_r = 0$	$M_r = 10^3$	$M_r = 10^4$	$M_r = 10^5$
$M = 10^3$	4.0×10^{-2}	2.6×10^{-2}	--	--
$M = 10^4$	1.2×10^{-2}	7.8×10^{-3}	2.3×10^{-3}	--
$M = 10^5$	3.9×10^{-3}	2.3×10^{-3}	7.9×10^{-4}	2.5×10^{-4}
$M = 10^6$	1.2×10^{-3}	8.2×10^{-4}	2.4×10^{-4}	7×10^{-5}

TABLE 3

Hard anharmonic potential. Two standard deviations of the estimator (4.33) in the case of potential (4.30) for different M and M_r . The other parameters are the same as in Table 2.

	$M_r = 0$	$M_r = 10^3$	$M_r = 10^4$	$M_r = 10^5$
$M = 10^3$	3.3×10^{-2}	2.3×10^{-2}	--	--
$M = 10^4$	1.1×10^{-2}	7.4×10^{-3}	3.0×10^{-3}	--
$M = 10^5$	3.5×10^{-3}	2.4×10^{-3}	9.5×10^{-4}	6.7×10^{-4}
$M = 10^6$	1.1×10^{-3}	7.4×10^{-4}	2.9×10^{-4}	2.2×10^{-4}

4.3. Pricing a binary asset-or-nothing call option. Consider the Black-Scholes equation for pricing a binary asset-or-nothing call option:

$$(4.34) \quad \frac{\partial u}{\partial t} + \frac{\nu^2}{2} x^2 \frac{\partial^2 u}{\partial x^2} + r x \frac{\partial u}{\partial x} - r u = 0, \quad 0 \leq t < T, \quad x \in \mathbf{R},$$

$$u(T, x) = f(x) = \begin{cases} 0 & \text{if } x < K, \\ x & \text{if } x \geq K. \end{cases}$$

The solution of this problem for $x > 0$ and $K > 0$ is

$$(4.35) \quad u(t, x) = x \Phi(y_*),$$

where

$$y_* = \frac{1}{\nu\sqrt{T-t}} \left[\ln \frac{x}{K} + \left(r + \frac{\nu^2}{2} \right) (T-t) \right] \quad \text{and} \quad \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-z^2/2} dz.$$

The probabilistic representation (with $\mu = 0$) of the solution to (4.34) takes the form

$$(4.36) \quad u(s, x) = E \left[f(X_{s,x}(T)) e^{-r(T-s)} + Z_{s,x}(T) \right],$$

$$(4.37) \quad dX = rX dt + \nu X dw(t), \quad X(s) = x,$$

$$(4.38) \quad dZ = F(t, X) e^{-r(t-s)} dw(t), \quad Z(s) = 0.$$

The purpose of this example is to illustrate that the approach to evaluating $u(s, x)$ introduced in section 3.2 works, in principle, in the case of discontinuous initial conditions $f(x)$. We use, as a set of basis functions for regression, the set consisting of three functions:

$$(4.39) \quad \varphi_1(x) = \frac{K}{\pi} (\arctan(\alpha(x-K)) + \arctan(\alpha K)),$$

$$\varphi_2(x) = \frac{x}{2} + \frac{x(x-2K)}{4(\sqrt{(x-K)^2/4 + \beta} + \sqrt{K^2/4 + \beta})}, \quad \varphi_3(x) = \frac{x}{\gamma + x^2},$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are parameters, which can change from one time layer to another. We note that the functions are chosen so that $\varphi_l(0) = 0$, $l = 1, 2, 3$, and the payoff $f(x)$ is well approximated by $\varphi_1(x) + \varphi_2(x)$ with large α and small β .

In the experiments, we take the volatility $\nu = 0.2$, the interest rate $r = 0.02$, and the maturity time $T = 3$ and approximate the option price $u(0, 1)$, whose exact value due to (4.35) is $u(0, 1) \approx 0.63548$. We define the time-dependent $\alpha = \alpha(t)$ and $\beta = \beta(t)$ via linear interpolation:

$$\alpha(t) = \frac{10t}{T} + \frac{0.01(T-t)}{T}, \quad \beta(t) = \frac{0.0001t}{T} + \frac{0.005(T-t)}{T},$$

and we choose $\gamma = 8$. We simulate (4.37)–(4.38) using the weak Euler scheme with time step $h = T/N = 0.001$. In the first run (see section 3.5 for the description of the algorithm), we put $F = 0$ and store the values $f(m\bar{X}(T))e^{-rT}$, which are needed for the regression in the second run. In the second run, using regression with the set of basis functions (4.39), we construct the estimator $\hat{u}(\theta_{k'}, x)$ for $u(\theta_{k'}, x)$, where $\theta_{k'} = \varkappa k' h$, $k' = 1, \dots, N'$; \varkappa and N' are nonnegative integers such that $\varkappa N' h = T$. We use here $\varkappa = 5$; i.e., we recalculate the estimator \hat{u} only once per five time layers to reduce the computational cost. Further, $\hat{u}(t_k, x)$ is set equal to zero for $0 \leq t_k < 0.01$. In the second run we put $F(t, x) = -\nu \partial \hat{u} / \partial x$. In both runs we simulate $M = 4 \cdot 10^4$ independent trajectories. As a result, we get $u(0, 1) \approx \bar{u}(0, 1) = 0.6358 \pm 0.0018$ with the fiducial probability 95%. To achieve a similar result without variance reduction, namely, $\bar{u}(0, 1) = 0.6342 \pm 0.0019$, one has to simulate $M = 5 \cdot 10^5$ independent trajectories, which requires at least three times more computational time than the procedure with variance reduction. This experiment demonstrates that the simple and cheap estimation of $\partial u / \partial x$ by $\partial \hat{u} / \partial x$ works even in the case of discontinuous initial conditions.

5. Conclusions. Starting an MC simulation, first of all we have to estimate the number of trajectories required to reach a prescribed accuracy. Fortunately, we can easily do this because a reliable estimate of the variance can be obtained by a preliminary numerical experiment using a relatively small set of trajectories. If the required number of trajectories is too large, we run inevitably into the problem of variance reduction. The known variance reduction methods (the method of importance sampling, the method of control variates, and the combining method) are based on the assumption that approximations of the solution $u(t, x)$ of the considered problem and its spatial derivatives $\partial u(t, x) / \partial x^i$ are known. In this paper we proposed to construct such approximations as a part of the MC simulation using conditional probabilistic representations together with the regression method and thus make the variance reduction methods practical. The basis used in the regression method can be chosen using some a priori knowledge of the considered problems, as illustrated in the examples.

As is known (see, e.g., [16]), the variance reduction methods are applicable in the case of boundary value problems for parabolic and elliptic equations as well. Although here we illustrated the proposed implementation of these variance reduction methods for the Cauchy problems for parabolic equations, the approach is straightforwardly applicable to boundary value problems.

We also note that the proposed technique of conditional probabilistic representations together with regression can be used for evaluating different Greeks for American- and Bermudan-type options (see [1]).

REFERENCES

- [1] D. BELOMESTNY, G. N. MILSTEIN, AND J. G. M. SCHOENMAKERS, *Sensitivities for Bermudan Options by Regression Methods*, WIAS preprint 1247, WIAS, Berlin, 2007.
- [2] B. BOUCHARD, I. EKELAND, AND N. TOUZI, *On the Malliavin approach to Monte Carlo approximation of conditional expectations*, Finance Stoch., 8 (2004), pp. 45–71.
- [3] J. FAN AND I. GIJBELS, *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London, 1996.
- [4] E. FOURNIÉ, J.-M. LASRY, J. LEBUCHOUX, AND P.-L. LIONS, *Application of Malliavin calculus to Monte Carlo methods in finance II*, Finance Stoch., 5 (2001), pp. 201–236.
- [5] E. FOURNIÉ, J.-M. LASRY, J. LEBUCHOUX, P.-L. LIONS, AND N. TOUZI, *Application of Malliavin calculus to Monte Carlo methods in finance*, Finance Stoch., 3 (1999), pp. 391–412.
- [6] S. A. GLADYSHEV AND G. N. MILSTEIN, *The Runge-Kutta method for calculation of Wiener integrals of functionals of exponential type*, Zh. Vychisl. Mat. i Mat. Fiz., 24 (1984), pp. 1136–1149.
- [7] P. GLASSERMAN, *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York, 2004.
- [8] L. GYÖRFI, M. KOHLER, A. KRZYŻAK, AND H. WALK, *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York, 2002.
- [9] A. KEBAIER, *Statistical Romberg extrapolation: A new variance reduction method and applications to option pricing*, Ann. Appl. Probab., 15 (2005), pp. 2681–2705.
- [10] A. KOHATSU-HIGA AND R. PETTERSSON, *Variance reduction methods for simulation of densities on Wiener space*, SIAM J. Numer. Anal., 40 (2002), pp. 431–450.
- [11] N. V. KRYLOV, *Controllable Processes of Diffusion Type*, Nauka, Moscow, 1977.
- [12] G. N. MILSTEIN, *Numerical Integration of Stochastic Differential Equations*, Ural State University, Sverdlovsk, 1988 (in Russian); English translation: Kluwer Academic, Dordrecht, The Netherlands, 1995.
- [13] G. N. MILSTEIN AND J. G. M. SCHOENMAKERS, *Monte Carlo construction of hedging strategies against multi-asset European claims*, Stoch. Stoch. Rep., 73 (2002), pp. 125–157.
- [14] G. N. MILSTEIN, J. G. M. SCHOENMAKERS, AND V. SPOKOINY, *Transition density estimation for stochastic differential equations via forward-reverse representations*, Bernoulli, 10 (2004), pp. 281–312.
- [15] G. N. MILSTEIN AND M. V. TRETYAKOV, *Quasi-symplectic methods for Langevin-type equations*, IMA J. Numer. Anal., 23 (2003), pp. 593–626.
- [16] G. N. MILSTEIN AND M. V. TRETYAKOV, *Stochastic Numerics for Mathematical Physics*, Springer-Verlag, Berlin, 2004.
- [17] G. N. MILSTEIN AND M. V. TRETYAKOV, *Numerical analysis of Monte Carlo evaluation of Greeks by finite differences*, J. Comput. Finance, 8 (2005), pp. 1–33.
- [18] G. N. MILSTEIN AND M. V. TRETYAKOV, *Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients*, SIAM J. Numer. Anal., 43 (2005), pp. 1139–1154.
- [19] G. N. MILSTEIN AND M. V. TRETYAKOV, *Computing ergodic limits for Langevin equations*, Phys. D, 229 (2007), pp. 81–95.
- [20] N. NEWTON, *Continuous-time Monte Carlo methods and variance reduction*, in Numerical Methods in Finance, L. C. G. Rodgers and D. Talay, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 22–42.
- [21] N. J. NEWTON, *Variance reduction for simulated diffusions*, SIAM J. Appl. Math., 54 (1994), pp. 1780–1805.
- [22] D. NUALART, *The Malliavin Calculus and Related Topics*, Springer-Verlag, Berlin, 2006.
- [23] R. REIGADA, A. H. ROMERO, A. SARMIENTO, AND K. LINDENBERG, *One-dimensional arrays of oscillators: Energy localization in thermal equilibrium*, J. Chem. Phys., 111 (1999), pp. 1373–1384.
- [24] W. WAGNER, *Monte Carlo evaluation of functionals of solutions of stochastic differential equations. Variance reduction and numerical examples*, Stoch. Anal. Appl., 6 (1988), pp. 447–468.
- [25] G. ZOU AND R. D. SKEEL, *Robust variance reduction for random walk methods*, SIAM J. Sci. Comput., 25 (2004), pp. 1964–1981.