Bernoulli **10**(2), 2004, 281–312

Transition density estimation for stochastic differential equations via forward–reverse representations

GRIGORI N. MILSTEIN,[1*,2] JOHN G.M. SCHOENMAKERS[1**] and VLADIMIR SPOKOINY[1†]

[1]*Weierstrass-Institut für Angewandte Analysis und Stochastik, Berlin, Germany.*
*E-mail: *Milstein@wiasberlin.de; **schoenma@wiasberlin.de; †spokoiny@wiasberlin.de*
[2]*Ural State University, Ekaterinburg, Russia*

The general reverse diffusion equations are derived and applied to the problem of transition density estimation of diffusion processes between two fixed states. For this problem we propose density estimation based on forward–reverse representations and show that this method allows essentially better results to be achieved than the usual kernel or projection estimation based on forward representations only.

Keywords: forward and reverse diffusion; Monte Carlo simulation; statistical estimation; transition density

1. Introduction

Consider the stochastic differential equation (SDE) in the Itô sense

$$\mathrm{d}X = a(s, X)\,\mathrm{d}s + \sigma(s, X)\,\mathrm{d}W(s), \qquad t_0 \leqslant s \leqslant T, \tag{1.1}$$

where $X = (X^1, \ldots, X^d)^{\mathrm{T}}$, $a = (a^1, \ldots, a^d)^{\mathrm{T}}$ are d-dimensional vectors, $W = (W^1, \ldots, W^m)^{\mathrm{T}}$ is an m-dimensional standard Wiener process, and $\sigma = \{\sigma^{ij}\}$ is a $\mathrm{d} \times m$ matrix, $m \geqslant d$. We assume that the $\mathrm{d} \times d$ matrix $b := \sigma\sigma^{\mathrm{T}}$, $b = \{b^{ij}\}$, is of full rank and, moreover, that the uniform ellipticity condition holds: there exists $\alpha > 0$ such that

$$\| \big(\sigma(s, x)\sigma^{\mathrm{T}}(s, x)\big)^{-1} \| \leqslant \alpha^{-1} \tag{1.2}$$

for all (s, x), $s \in [t_0, T]$, $x \in \mathbb{R}^d$ and some $\alpha > 0$. The functions $a^i(s, x)$ and $\sigma^{ij}(s, x)$ are assumed to satisfy the same regularity conditions as in Bally and Talay (1996b), that is, their derivatives of any order exist and are bounded. In particular, this implies existence and uniqueness of the solution $X_{t,x}(s) \in \mathbb{R}^d$, $X_{t,x}(t) = x$, $t_0 \leqslant t \leqslant s \leqslant T$, of (1.1), smoothness of the transition density $p(t, x, s, y)$ of the Markov process X, and existence of exponential bounds for the density and its derivatives with respect to $t > t_0$, x, y.

The aim of this paper is the construction of a Monte Carlo estimator of the unknown transition density $p(t, x, T, y)$ for fixed t, x, T, y, which improves upon classical kernel or projection estimators based on realizations of $X_{t,x}(T)$ directly.

Classical Monte Carlo methods allow for effective estimation of functionals of the form

$$I(f) = \int p(t, x, T, y) f(y)\, \mathrm{d}y \tag{1.3}$$

for smooth functions f not increasing too rapidly. These methods exploit the probabilistic representation $I(f) = \mathrm{E}\, f(X_{t,x}(T))$. Let $\overline{X}_{t,x}$ be an approximation of the process $X_{t,x}$ and let $\overline{X}^{(n)}_{t,x}(T)$, for $n = 1, \ldots, N$, be independent realizations of $\overline{X}_{t,x}(T)$. Then, provided the accuracy of the approximation of $X_{t,x}$ by $\overline{X}_{t,x}$ is sufficiently good, $I(f)$ may be estimated by

$$\hat{\bar{I}} = \frac{1}{N} \sum_{n=1}^{N} f\left(\overline{X}^{(n)}_{t,x}(T)\right)$$

with root-N accuracy, that is a statistical error of order $N^{-1/2}$.

The problem of estimating the transition density of a diffusion process is more involved; see Bally and Talay (1996a), Hu and Watanabe (1996) and Kohatsu-Higa (1997). For an approximation $\overline{X}_{t,x}$, it is natural to expect that its transition density $\bar{p}(t, x, T, y)$ is an approximation of $p(t, x, T, y)$. Indeed, if $\overline{X}_{t,x}(T, h)$ is the approximation of $X_{t,x}(T)$ obtained via numerical integration by the strong Euler scheme with time step h, then the density $\bar{p}_h(t, x, T, y)$ converges to $p(t, x, T, y)$ uniformly in y when the step size h tends to zero. More precisely:

$$p(t, x, T, y) - \bar{p}_h(t, x, T, y) = hC(t, x, T, y) + h^2 R_h(t, x, T, y), \tag{1.4}$$

with

$$|C(t, x, T, y)| + |R_h(t, x, T, y)| \leqslant \frac{K}{(T-t)^q} \exp\left(-c\frac{|x-y|^2}{T-t}\right),$$

where K, c, q are some positive constants; see Bally and Talay (1996b). Strictly speaking, (1.4) is derived in Bally and Talay (1996b) for autonomous systems. However, there is no doubt that under our assumptions of smoothness, boundedness and uniform ellipticity this result holds for the non-autonomous case as well.

Further, Hu and Watanabe (1996) and Kohatsu-Higa (1997) show that the quantity

$$\tilde{p}_h(t, x, T, y) = \mathrm{E}\, \phi_h(\overline{X}_{t,x}(T, h) - y),$$

with $\phi_h(x) = (2\pi h^2)^{-\mathrm{d}/2} \exp\{-|x|^2/(2h^2)\}$, converges to $p(t, x, T, y)$ as $h \to 0$. Hu and Watanabe (1996) used schemes of numerical integration in the strong sense, while Kohatsu-Higa (1997) applied numerical schemes in a weak sense. Combining this result with the classical Monte Carlo methods leads to the following estimator of the transition density:

$$\hat{\tilde{p}}(t, x, T, y) = \frac{1}{N} \sum_{n=1}^{N} \phi_h\left(\overline{X}_n - y\right), \tag{1.5}$$

where $\overline{X}_n = \overline{X}^{(n)}_{t,x}(T, h)$, $n = 1, \ldots, N$, are independent realizations of $\overline{X}_{t,x}(T, h)$.

More generally, one may estimate the transition density $p(t, x, T, y)$ from the sample $X_n = X^{(n)}_{t,x}(T)$ by using standard methods of nonparametric statistics. For example, the

kernel (Parzen–Rosenblatt) density estimator with a kernel K and a bandwidth δ is given by

$$\hat{p}(t, x, T, y) = \frac{1}{N\delta^d} \sum_{n=1}^{N} K\left(\frac{X_n - y}{\delta}\right); \tag{1.6}$$

see, for example, Devroye and Györfi (1985), Silverman (1986) or Scott (1992). Of course, in reality we have only the approximation $\overline{X}_n$ instead of X_n and so we obtain the estimator

$$\hat{\bar{p}}_h(t, x, T, y) = \frac{1}{N\delta^d} \sum_{n=1}^{N} K\left(\frac{\overline{X}_n - y}{\delta}\right). \tag{1.7}$$

Clearly, proposal (1.5) is a special case of estimator (1.7) with kernel K being the standard normal density and bandwidth δ equal to the step of numerical integration h.

The estimation loss $\hat{\bar{p}}_h(t, x, T, y) - p(t, x, T, y)$ can be split into an error $\hat{\bar{p}}_h - \hat{p}$ due to numerical approximation of the process X by $\overline{X}$ and an error $\hat{p} - p$ due to the kernel estimation which depends on the sample size N, the bandwidth δ and the kernel K. The loss of the first kind can be reduced considerably by properly selecting a scheme of numerical integration and choosing a small step h. The more important loss, however, is caused by the kernel estimation. It is well known that the quality of density estimation strongly depends on the bandwidth δ and the choice of a suitable bandwidth is a delicate issue (see Devroye and Györfi 1985). Even an optimal choice of the bandwidth δ leads to quite poor estimation quality, particularly for large dimension d. More specifically, if the underlying density is known to be twice continuously differentiable then the optimal bandwidth δ is of order $N^{-1/(4+d)}$, leading to accuracy of order $N^{-2/(4+d)}$; see Bretagnolle and Carol-Huber (1979), Scott (1992) or Silverman (1986). For larger d, this would require a huge sample size N to provide reasonable accuracy of estimation. In the statistical literature this problem is referred to as the 'curse of dimensionality'.

In this paper we propose a method of density estimation which is generally root-N consistent and thus avoids the curse of dimensionality. In Section 2 we consider probabilistic representations for the functionals $I(f)$ in (1.3), which provide different Monte Carlo methods for the evaluation of $I(f)$. We also show how the variance of the Monte Carlo estimation can be reduced by the choice of a suitable probabilistic representation. Then, in Section 3, we introduce the reverse diffusion process in order to derive probabilistic representations for functionals of the form

$$I^*(g) = \int g(x) p(t, x, T, y)\, \mathrm{d}x. \tag{1.8}$$

Clearly, the 'curse of dimensionality' is not encountered in the estimation of functionals $I(f)$ in (1.3) using forward representations. Similarly, as we shall see in Section 3, Monte Carlo estimation of functionals of the form (1.8) via probabilistic representations based on reverse diffusion can also be carried out with root-N accuracy. These important features have been utilized in the central theme of this paper, the development of a new method for estimating the transition density $p(t, x, T, y)$ of a diffusion process which generally allows for root-N consistent estimation for prespecified values of t, x, T and y (we emphasize that

the problem of estimating $p(t, x, T, y)$ for fixed t, x, T and y is more difficult than the problem of estimating the integrals $I(f)$, $I(f, g)$ or $I^*(g)$). This method, which is presented in Section 4, is based on a combination of forward representation (1.3) and reverse representation (1.8) via the Chapman–Kolmogorov equation and has led to two different types of estimators called *kernel* and *projection* estimators. General properties of these estimators are studied in Sections 6 and 7. Before that, in Section 5, we demonstrate the advantages of combining the forward and reverse diffusion for transition density estimation in a simple one-dimensional example. We show by an explicit analysis of an Ornstein–Uhlenbeck type process that root-N accuracy can be achieved.

Throughout Sections 5–7 all results are derived with respect to exact solutions of the respective SDEs. In Section 8 we study in particular the estimation loss due to application of the strong Euler scheme with discretization step h of various kernel estimators and find that this loss is of order $O(h)$, uniform in the bandwidth δ.

In Section 9 we compare the computational complexity of the forward–reverse estimators with pure forward estimators and give some numerical results for the example in Section 5. We conclude that, in general, for the problem of estimating the transition density between two particular states the forward–reverse estimator outperforms the usual estimator based on forward diffusion only.

2. Probabilistic representations based on forward diffusion

In this section we present a general probabilistic representation and the corresponding Monte Carlo estimator for a functional of the form (1.3). We also show that the variance of the Monte Carlo method can be reduced by choosing a proper representation.

For a given function f, the function

$$u(t, x) = \mathrm{E}\, f(X_{t,x}(T)) = \int p(t, x, T, y) f(y)\, \mathrm{d}y \tag{2.1}$$

is the solution of the Cauchy problem for the parabolic equation

$$Lu := \frac{\partial u}{\partial t} + \frac{1}{2}\sum_{i,j=1}^{d} b^{ij}(t, x)\frac{\partial^2 u}{\partial x^i \partial x^j} + \sum_{i=1}^{d} a^i(t, x)\frac{\partial u}{\partial x^i} = 0, \qquad u(T, x) = f(x).$$

Via the probabilistic representation (2.1), $u(t, x)$ may be computed by Monte Carlo simulation using weak methods for numerical integration of SDE (1.1). Let $\overline{X}$ be an approximation of the process X in (1.1), obtained by some numerical integration scheme. With $\overline{X}_{t,x}^{(n)}(T)$ being independent realizations of $\overline{X}_{t,x}(T)$, the value $u(t, x)$ can be estimated by

$$\hat{\hat{\boldsymbol{u}}} = \frac{1}{N}\sum_{n=1}^{N} f\left(\overline{X}_{t,x}^{(n)}(T)\right). \tag{2.2}$$

Moreover, by taking a random initial value $X(t) = \xi$, where the random variable ξ has a density g, we obtain a probabilistic representation for integrals of the form

$$I(f, g) = \iint g(x)p(t, x, T, y)f(y)\,\mathrm{d}x\,\mathrm{d}y. \tag{2.3}$$

The estimation error $|\hat{\bar{u}} - u|$ of the estimator $\hat{\bar{u}}$ in (2.2) is due to the Monte Carlo method and to the numerical integration of SDE (1.1). The second error can be reduced by selecting a suitable method and step of numerical integration. The first one, the Monte Carlo error, is of order $\{N^{-1}\operatorname{var} f(\bar{X}_{t,x}(T))\}^{1/2} \simeq \{N^{-1}\operatorname{var} f(X_{t,x}(T))\}^{1/2}$ and can, in general, be reduced by using variance reduction methods. Variance reduction methods can be derived from the following generalized probabilistic representation for $u(t, x)$:

$$u(t, x) = \mathrm{E}\,[\,f(X_{t,x}(T))\mathcal{X}_{t,x}(T) + \mathbb{X}_{t,x}(T)], \tag{2.4}$$

where $X_{t,x}(s)$, $\mathcal{X}_{t,x}(s)$, $\mathbb{X}_{t,x}(s)$, $s \geqslant t$, is the solution of the system of SDEs given by

$$\begin{aligned} &\mathrm{d}X = (a(s, X) - \sigma(s, X)h(s, X))\,\mathrm{d}s + \sigma(s, X)\,\mathrm{d}W(s), && X(t) = x,\\ &\mathrm{d}\mathcal{X} = h^{\mathrm{T}}(s, X)\mathcal{X}\,\mathrm{d}W(s), && \mathcal{X}(t) = 1,\\ &\mathrm{d}\mathbb{X} = F^{\mathrm{T}}(s, X)\mathcal{X}\,\mathrm{d}W(s), && \mathbb{X}(t) = 0. \end{aligned} \tag{2.5}$$

In (2.5), $\mathcal{X}$ and $\mathbb{X}$ are scalars, and $h(t, x) = (h^1(t, x), \ldots, h^m(t, x))^{\mathrm{T}} \in \mathbb{R}^m$ and $F(t, x) = (F^1(t, x), \ldots, F^m(t, x))^{\mathrm{T}} \in \mathbb{R}^m$ are vector functions satisfying some regularity conditions (for example, they are sufficiently smooth and have bounded derivatives). The usual probabilistic representation (2.1) is a particular case of (2.4)–(2.5) with $h = 0$, $F = 0$; see, for example, Dynkin (1965). The representation for $h \neq 0$, $F = 0$ follows from Girsanov's theorem and then we obtain (2.4) since $\mathrm{E}\,\mathbb{X} = 0$.

Consider the random variable $\eta := f(X_{t,x}(T))\mathcal{X}_{t,x}(T) + \mathbb{X}_{t,x}(T)$. While the mathematical expectation $\mathrm{E}\,\eta$ does not depend on h and F, the variance $\operatorname{var}\eta = \mathrm{E}\,\eta^2 - (\mathrm{E}\,\eta)^2$ does. The Monte Carlo error in the estimation of (2.4) is of order $\sqrt{N^{-1}\operatorname{var}\eta}$ and so by reduction of the variance $\operatorname{var}\eta$ the Monte Carlo error may be reduced. Two variance reduction methods are well known: the method of importance sampling where $F = 0$ (see Milstein 1995; Newton 1994; Wagner 1988), and the method of control variates where $h = 0$ (see Newton 1994). For both methods it is shown that for a sufficiently smooth function f the variance can be reduced to zero. A more general statement by Milstein and Schoenmakers (2002) is given in Theorem 2.1 below. Introduce the process

$$\eta(s) = u(s, X_{t,x}(s))\mathcal{X}_{t,x}(s) + \mathbb{X}_{t,x}(s), \qquad t \leqslant s \leqslant T.$$

Clearly $\eta(t) = u(t, x)$ and $\eta(T) = f(X_{t,x}(T))\mathcal{X}_{t,x}(T) + \mathbb{X}_{t,x}(T)$.

Theorem 2.1. *Let h and F be such that for any $x \in \mathbb{R}^d$ there is a solution to the system* (2.5) *on the interval* $[t, T]$. *Then the variance* $\operatorname{var}\eta(T)$ *is equal to*

$$\operatorname{var}\eta(T) = \mathrm{E}\int_t^T \mathcal{X}_{t,x}^2(s)\sum_{j=1}^m\left(\sum_{i=1}^d \sigma^{ij}\frac{\partial u}{\partial x^i} + uh^j + F^j\right)^2 \mathrm{d}s \tag{2.6}$$

provided that the mathematical expectation in (2.6) *exists.*

In particular, if h and F satisfy

$$\sum_{i=1}^{d} \sigma^{ij} \frac{\partial u}{\partial x^i} + uh^j + F^j = 0, \qquad j = 1, \ldots, m,$$

then $\operatorname{var} \eta(T) = 0$ *and* $\eta(s)$ *is deterministic and independent of* $s \in [t, T]$.

Proof. The Itô formula implies

$$\mathrm{d}\eta(s) = \mathcal{X}_{t,x}(s)(Lu)\,\mathrm{d}s + \mathcal{X}_{t,x}(s) \sum_{j=1}^{m} \left(\sum_{i=1}^{d} \sigma^{ij} \frac{\partial u}{\partial x^i} + uh^j + F^j \right) \mathrm{d}W^j(s),$$

and then by $Lu = 0$ we have

$$\eta(s) = \eta(t) + \int_t^s \mathcal{X}_{t,x}(s') \sum_{j=1}^{m} \left(\sum_{i=1}^{d} \sigma^{ij} \frac{\partial u}{\partial x^i} + uh^j + F^j \right) \mathrm{d}W^j(s').$$

Hence, (2.6) follows and the last assertion is obvious. □

Remark 2.1. Clearly, h and F in Theorem 2.1 cannot be constructed without knowing $u(s, x)$. Nevertheless, the theorem claims a general possibility of variance reduction by proper choice of the functions h^j and F^j, $j = 1, \ldots, m$.

3. Representations relying on reverse diffusion

In the previous section a broad class of probabilistic representations for the integral functionals $I(f) = \int f(y) p(t, x, T, y)\,\mathrm{d}y$, and more generally for the functionals $I(f, g) = \int\int g(x) p(t, y, T, y) f(y)\,\mathrm{d}x\,\mathrm{d}y$, is described. Another approach is based on the so-called *reverse diffusion* and was introduced by Thomson (1987) (see also Kurbanmuradov *et al.* 1999; 2001). In this section we derive the reverse diffusion system in a more transparent and more rigorous way. The method of reverse diffusion provides a probabilistic representation (hence a Monte Carlo method) for functionals of the form

$$I^*(g) = \int g(x) p(t, x, T, y)\,\mathrm{d}x, \tag{3.1}$$

where g is a given function. This representation may easily be extended to the functionals $I(f, g)$ from (2.3).

For a given function g and fixed t we define

$$v(s, y) := \int g(x') p(t, x', s, y)\,\mathrm{d}x', \qquad s > t,$$

and consider the Fokker–Planck equation (forward Kolmogorov equation) for $p(t, x, s, y)$,

$$\frac{\partial p}{\partial s} = \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial y^i \partial y^j}(b^{ij}(s,\, y)p) - \sum_{i=1}^{d} \frac{\partial}{\partial y^i}(a^i(s,\, y)p).$$

Then, multiplying this equation by $g(x)$ and integrating with respect to x yields the following Cauchy problem for the function $v(s,\, y)$:

$$\frac{\partial v}{\partial s} = \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial y^i \partial y^j}(b^{ij}(s,\, y)v) - \sum_{i=1}^{d} \frac{\partial}{\partial y^i}(a^i(s,\, y)v), \qquad s > t,$$
$$v(t,\, y) = g(y).$$

We introduce the reversed time variable $\tilde{s} = T + t - s$ and define

$$\tilde{v}(\tilde{s},\, y) = v(T + t - \tilde{s},\, y),$$

$$\tilde{a}^i(\tilde{s},\, y) = a^i(T + t - \tilde{s},\, y),$$

$$\tilde{b}^{ij}(\tilde{s},\, y) = b^{ij}(T + t - \tilde{s},\, y).$$

Clearly, $v(T,\, y) = \tilde{v}(t,\, y)$ and

$$\frac{\partial \tilde{v}}{\partial \tilde{s}} + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial y^i \partial y^j}(\tilde{b}^{ij}(\tilde{s},\, y)\tilde{v}) - \sum_{i=1}^{d} \frac{\partial}{\partial y^i}(\tilde{a}^i(\tilde{s},\, y)\tilde{v}) = 0, \qquad \tilde{s} < T,$$
$$\tilde{v}(T,\, y) = v(t,\, y) = g(y). \tag{3.2}$$

Since $b^{ij} = b^{ji}$ and so $\tilde{b}^{ij} = \tilde{b}^{ji}$, the partial differential equation in (3.2) may be written in the form (with s instead of $\tilde{s}$)

$$\tilde{L}\tilde{v} := \frac{\partial \tilde{v}}{\partial s} + \frac{1}{2}\sum_{i,j=1}^{d} \tilde{b}^{ij}(s,\, y)\frac{\partial^2 \tilde{v}}{\partial y^i \partial y^j} + \sum_{i=1}^{d} \alpha^i(s,\, y)\frac{\partial \tilde{v}}{\partial y^i} + c(s,\, y)\tilde{v} = 0, \qquad s < T, \tag{3.3}$$

where

$$\alpha^i(s,\, y) = \sum_{j=1}^{d} \frac{\partial \tilde{b}^{ij}}{\partial y^j} - \tilde{a}^i, \qquad c(s,\, y) = \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 \tilde{b}^{ij}}{\partial y^i \partial y^j} - \sum_{i=1}^{d} \frac{\partial \tilde{a}^i}{\partial y^i}. \tag{3.4}$$

We thus obtain a Cauchy problem in reverse time and may state the following result.

Theorem 3.1. *$I^*(g)$ has a probabilistic representation*

$$I^*(g) = v(T,\, y) = \tilde{v}(t,\, y) = \mathrm{E}[g(Y_{t,y}(T))\mathcal{Y}_{t,y}(T)], \tag{3.5}$$

where the vector process $Y_{t,y}(s) \in \mathbb{R}^d$ and the scalar process $\mathcal{Y}_{t,y}(s)$ solve the stochastic system

$$\begin{aligned} \mathrm{d}Y &= \alpha(s,\, Y)\,\mathrm{d}s + \tilde{\sigma}(s,\, Y)\,\mathrm{d}\tilde{W}(s), & Y(t) &= y, \\ d\mathcal{Y} &= c(s,\, Y)\mathcal{Y}\,\mathrm{d}s, & \mathcal{Y}(t) &= 1, \end{aligned} \tag{3.6}$$

with $\tilde{\sigma}(s,\, y) = \sigma(T + t - s,\, y)$ and $\tilde{W}$ being an m-dimensional standard Wiener process.

It is natural to call (3.6) *the reverse system* of (1.1). The probabilistic representation (3.5)–(3.6) for the integral (3.1) leads naturally to the Monte Carlo estimator $\hat{\boldsymbol{v}}$ for $v(T, y)$,

$$\hat{\boldsymbol{v}} = \frac{1}{M}\sum_{m=1}^{M} g\Big(\overline{Y}_{t,y}^{(m)}(T)\Big)\overline{\mathcal{Y}}_{t,y}^{(m)}(T), \tag{3.7}$$

where $(\overline{Y}_{t,y}^{(m)}, \overline{\mathcal{Y}}_{t,y}^{(m)})$, $m = 1, \ldots, M$, are independent realizations of the process $(\overline{Y}_{t,y}, \overline{\mathcal{Y}}_{t,y})$ that approximates the process $(Y_{t,y}, \mathcal{Y}_{t,y})$ from (3.6).

Similarly to (2.4)–(2.5), the representation (3.5)–(3.6) may be extended to

$$v(T, y) = \mathrm{E}\,[g(Y_{t,y}(T))\mathcal{Y}_{t,y}(T) + \mathbb{Y}_{t,y}(T)], \tag{3.8}$$

where $Y_{t,y}(s)$, $\mathcal{Y}_{t,y}(s)$, $\mathbb{Y}_{t,y}(s)$, $s \geqslant t$, solve the following system of SDEs:

$$\begin{array}{ll} \mathrm{d}Y = (\alpha(s, Y) - \tilde{\sigma}(s, Y)\tilde{h}(s, Y))\,\mathrm{d}s + \tilde{\sigma}(s, Y)\,\mathrm{d}\tilde{W}(s), & Y(t) = y, \\ d\mathcal{Y} = c(s, Y)\mathcal{Y}\,\mathrm{d}s + \tilde{h}^{\mathrm{T}}(s, Y)\mathcal{Y}\,\mathrm{d}\tilde{W}(s), & \mathcal{Y}(t) = 1, \\ \mathrm{d}\mathbb{Y} = \tilde{F}^{\mathrm{T}}(s, Y)\mathcal{Y}\,\mathrm{d}\tilde{W}(s), & \mathbb{Y}(t) = 0. \end{array} \tag{3.9}$$

In (3.9), $\mathcal{Y}$ and $\mathbb{Y}$ are scalars and $\tilde{h}(t, x) \in \mathbb{R}^m$ and $\tilde{F}(t, x) \in \mathbb{R}^m$ are arbitrary vector functions which satisfy some regularity conditions.

Remark 3.1. If system (1.1) is autonomous, then $\tilde{b}^{ij}$, $\tilde{a}^i$, α^i, $\tilde{\sigma}$, and c depend on y only, $\tilde{b}^{ij}(y) = b^{ij}(y)$, $\tilde{a}^i(y) = a^i(y)$, and so $\tilde{\sigma}(y)$ can be taken equal to $\sigma(y)$.

Remark 3.2. By constructing the reverse system of reverse system (3.6), we obtain the original system (1.1) accompanied by a scalar equation with coefficient $-c$. By then taking the reverse of this system we obtain (3.6) again.

Remark 3.3. If the original stochastic system (1.1) is linear, then the system (3.6) is linear as well and c depends on t only.

Remark 3.4. variance reduction methods discussed in Section 2 may be applied to the reverse system as well. In particular, for the reverse system a theorem analogue to Theorem 2.1 applies.

4. Transition density estimation based on forward–reverse representations

In this section we present estimators for the target probability density $p(t, x, T, y)$, which utilize both the forward and the reverse diffusion system. More specifically, we give two different Monte Carlo estimators for $p(t, x, T, y)$ based on forward–reverse representations: a forward–reverse kernel estimator and a forward–reverse projection estimator. A detailed analysis of the performance of these estimators is postponed to Sections 6 and 7.

We start with a heuristic discussion. Let t^* be an internal point of the interval $[t, T]$. By the Kolmogorov–Chapman equation for the transition density we have

$$p(t, x, T, y) = \int p(t, x, t^*, x')\, p(t^*, x', T, y)\, \mathrm{d}x'. \tag{4.1}$$

By applying Theorem 3.1 with $g(x') = p(t, x, t^*, x')$, it follows that this equation has a probabilistic representation

$$p(t, x, T, y) = \mathrm{E}\, p(t, x, t^*, Y_{t^*,y}(T))\, \mathcal{Y}_{t^*,y}(T). \tag{4.2}$$

Since in general the density function $x' \to p(t, x, t^*, x')$ is also unknown, we cannot apply the Monte Carlo estimator $\hat{\boldsymbol{v}}$ in (3.7) to representation (4.2) directly. However, the key idea is now to estimate this density function from a sample of independent realizations of X on the interval $[t, t^*]$ by standard methods of nonparametric statistics and then to replace the unknown density function in the right-hand side of (4.2) by its estimator, say $x' \to \hat{p}(t, x, t^*, x')$. This idea suggests the following procedure. Generate by numerical integration of the forward system (1.1) and the reverse system (3.6) (or (3.9)) independent samples $\overline{X}_{t,x}^{(n)}(t^*)$, $n = 1, \ldots, N$, and $(\overline{Y}_{t^*,y}^{(m)}(T), \overline{\mathcal{Y}}_{t^*,y}^{(m)}(T))$, $m = 1, \ldots, M$, respectively (in general, different step sizes may be used for $\overline{X}$ and $\overline{Y}$). Let $\hat{\bar{p}}(t, x, t^*, x')$ be, for instance, the kernel estimator of $p(t, x, t^*, x')$ from (1.7), that is,

$$\hat{\bar{p}}(t, x, t^*, x') = \frac{1}{N\delta^d} \sum_{n=1}^{N} K\left(\frac{\overline{X}_{t,x}^{(n)}(t^*) - x'}{\delta}\right).$$

Thus, replacing p by this kernel estimator in the right-hand side of reverse representation (4.2) yields a representation which may be estimated by

$$\hat{\bar{p}}(t, x, T, y) = \frac{1}{M}\left[\frac{1}{N\delta^d} \sum_{m=1}^{M} \sum_{n=1}^{N} K\left(\frac{\overline{X}_{t,x}^{(n)}(t^*) - \overline{Y}_{t^*,y}^{(m)}(T)}{\delta}\right) \overline{\mathcal{Y}}_{t^*,y}^{(m)}(T)\right]. \tag{4.3}$$

The estimator (4.3) will be called a forward–reverse kernel estimator.

We will show that the above heuristic idea does work and leads to estimators which have superior properties in comparison with the usual density estimators based on pure forward or pure reverse representations. Of course, the kernel estimation of $p(t, x, t^*, x')$ in the first step will as usual be crude for a particular x'. But, due to a good overall property of kernel estimators – the fact that any kernel estimator is a density – the impact of these pointwise errors will be reduced in the second step, the estimation of (4.2). In fact, by the Chapman–Kolmogorov equation (4.1) the estimation of the density at one point is done via the estimation of a functional of the form (4.2). It can be seen that the latter estimation problem has smaller degree of ill-posedness, and therefore the accuracy achievable for a given amount of computational effort will be improved.

Now we proceed with a formal description which essentially utilizes the next general result naturally extending Theorem 3.1.

Theorem 4.1. *For a bivariate function f we have*

$$J(f) := \iint p(t, x, t^*, x')p(t^*, y', T, y)f(x', y')\,\mathrm{d}x'\,\mathrm{d}y'$$

$$= \mathrm{E}\,[\,f(X_{t,x}(t^*), Y_{t^*,y}(T))\,\mathcal{Y}_{t^*,y}(T)], \tag{4.4}$$

where $X_{t,x}(s)$ obeys the forward equation (1.1) *and $(Y_{t^*,y}(s), \mathcal{Y}_{t^*,y}(s))$, $s \geqslant t^*$, is the solution of the reverse system* (3.6).

Proof. Conditioning on $X_{t,x}(t^*)$ and applying Theorem 3.1 with $g(\cdot) = f(x', \cdot)$ for every x' yields

$$\mathrm{E}\left(f(X_{t,x}(t^*), Y_{t^*,y}(T))\,\mathcal{Y}_{t^*,y}(T)\right) = \mathrm{E}\,\mathrm{E}\left(f(X_{t,x}(t^*), Y_{t^*,y}(T))\,\mathcal{Y}_{t^*,y}(T)|X_{t,x}(t^*)\right)$$

$$= \int p(t, x, t^*, x')\left(\int f(x', y')p(t^*, y', T, y)\,\mathrm{d}y'\right)\mathrm{d}x'.$$

□

Let $\overline{X}^{(n)}_{t,x}(t^*)$, $n = 1, \ldots, N$, be a sample of independent realizations of an approximation $\overline{X}$ of X, obtained by numerical integration of (1.1) on the interval $[t, t^*]$. Similarly, let $(\overline{Y}^{(m)}_{t^*,y}(T)\overline{\mathcal{Y}}^{(m)}_{t^*,y}(T))$, $m = 1, \ldots, M$, be independent realizations of a numerical solution of (3.6) on the interval $[t^*, T]$. Then the representation (4.4) leads to the following Monte Carlo estimator for $J(f)$:

$$\hat{\hat{\boldsymbol{J}}} = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M} f\left(\overline{X}^{(n)}_{t,x}(t^*), \overline{Y}^{(m)}_{t^*,y}(T)\right)\overline{\mathcal{Y}}^{(m)}_{t^*,y}(T). \tag{4.5}$$

Formally, $J(f) \to p(t, x, T, y)$ as $f \to \delta_{\mathrm{diag}}$ (in the distribution sense), where $\delta_{\mathrm{diag}}(x', y') := \delta_0(x' - y')$ and δ_0 is the Dirac function concentrated at zero. So, in the attempt to estimate the density $p(t, x, T, y)$, two families of functions f naturally arise. Let us take functions f of the form

$$f(x', y') =: f_{K,\delta}(x', y') = \delta^{-d}K\left(\frac{x' - y'}{\delta}\right),$$

where $\delta^{-d}K(u/\delta)$ converge to $\delta_0(u)$ (in the distribution sense) as $\delta \downarrow 0$. Then the corresponding expression for $\hat{J}$ coincides with the forward–reverse kernel estimator $\hat{p}$ in (4.3). As an alternative, consider functions f of the form

$$f(x', y') =: f_{\varphi,L}(x', y') = \sum_{\ell=1}^{L}\varphi_\ell(x')\varphi_\ell(y'),$$

where $\{\varphi_\ell,\ \ell \geqslant 1\}$ is a total orthonormal system in the function space $L_2(\mathbb{R}^d)$ and L is a natural number. It is known that $f_{\varphi,L} \to \delta_{\mathrm{diag}}$ (in the distribution sense) as $L \to \infty$. This leads to the forward–reverse projection estimator,

$$\hat{\bar{p}}_{pr} = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{\ell=1}^{L}\varphi_\ell\left(\overline{X}_{t,x}^{(n)}(t^*)\right)\varphi_\ell\left(\overline{Y}_{t^*,y}^{(m)}(T)\right)\overline{\mathcal{Y}}_{t^*,y}^{(m)}(T) = \sum_{\ell=1}^{L}\hat{\bar{\alpha}}_\ell\hat{\bar{\gamma}}_\ell, \tag{4.6}$$

with

$$\hat{\bar{\alpha}}_\ell = \frac{1}{N}\sum_{n=1}^{N}\varphi_\ell\left(\overline{X}_{t,x}^{(n)}(t^*)\right), \qquad \hat{\bar{\gamma}}_\ell = \frac{1}{M}\sum_{m=1}^{M}\varphi_\ell\left(\overline{Y}_{t^*,y}^{(m)}(T)\right)\overline{\mathcal{Y}}_{t^*,y}^{(m)}(T).$$

The general properties of the forward–reverse kernel estimator are studied in Section 6 and the forward–reverse projection estimator is studied in Section 7. As mentioned previously, by properly selecting a numerical integration scheme and step size h, approximate solutions of systems of SDEs can be simulated sufficiently close to exact solutions. Therefore, in Sections 6 and 7 the analysis is carried out with respect to exact solutions $X_{t,x}(s)$ and $(Y_{t^*,y}(s), \mathcal{Y}_{t^*,y}(s))$. For the impact of their approximations $\overline{X}_{t,x}(s)$ and $(\overline{Y}_{t^*,y}(s), \overline{\mathcal{Y}}_{t^*,y}(s))$ obtained by the Euler scheme on the estimation accuracy, we refer to Section 8.

Remark 4.1. If we take $t^* = T$ in (4.3) we obtain the usual forward kernel estimator (1.6) again. Indeed, for $t^* = T$ we have $\overline{Y}_{T,y}^{(m)}(T) = y$ and $\overline{\mathcal{Y}}_{T,y}^{(m)}(T) = 1$ for any m. Similarly, taking $t^* = t$ in (4.3) leads to the pure reverse estimator,

$$\hat{\bar{p}}(t, x, T, y) := \frac{1}{M\delta^d}\sum_{m=1}^{M}K\left(\frac{x - \overline{Y}_{t,y}^{(m)}(T)}{\delta}\right)\overline{\mathcal{Y}}_{t,y}^{(m)}(T). \tag{4.7}$$

It should be noted that the pure forward estimator gives for fixed x and one simulation sample of $\overline{X}$ an estimation of the density $p(t, x, T, y)$ for all y. On the other hand, the pure reverse estimator gives for fixed y and one simulation of the reverse system a density estimation for all x. In contrast, the proposed forward–reverse estimators require for each pair (x, y) a simulation of both the forward and the reverse process. However, as we will see, these estimators have superior convergence properties.

Remark 4.2. In general, it is possible to apply variance reduction methods to the estimator $\hat{J}$ in (4.5), based on the extended representations (2.4)–(2.5) and (3.8)–(3.9).

5. Explicit analysis of the forward–reverse kernel estimator in a one-dimensional example

We consider an example of a one-dimensional diffusion for which all characteristics of the forward–reverse kernel estimator introduced in Section 4 can be derived analytically. For constant a, b, the one-dimensional diffusion is given by the SDE

$$\mathrm{d}X = aX\,\mathrm{d}t + b\,\mathrm{d}W(t), \qquad X(0) = x, \tag{5.1}$$

which, for $a < 0$, is known as the Ornstein–Uhlenbeck process. By (3.6), the reverse system belonging to (5.1) is given by

$$\mathrm{d}Y = -aY\,\mathrm{d}s + b\,\mathrm{d}\tilde{W}(s), \qquad Y(t) = y,\ s > t, \tag{5.2}$$

$$\mathrm{d}\mathcal{Y} = -a\mathcal{Y}\,\mathrm{d}s, \qquad \mathcal{Y}(t) = 1. \tag{5.3}$$

Both systems (5.1) and (5.2) can be solved explicitly. Their solutions are given by

$$X(t) = \mathrm{e}^{at}\left(x + b\int_0^t \mathrm{e}^{-au}\,\mathrm{d}W(u)\right)$$

and

$$Y(s) = \mathrm{e}^{-a(s-t)}\left(y + b\int_t^s \mathrm{e}^{a(u-t)}\,\mathrm{d}\tilde{W}(u)\right),$$

$$\mathcal{Y}(s) = \mathrm{e}^{-a(s-t)},$$

respectively. It follows that

$$\mathrm{E}\,X(t) = \mathrm{e}^{at}x, \quad \operatorname{var} X(t) = b^2\mathrm{e}^{2at}\int_0^t \mathrm{e}^{-2au}\,\mathrm{d}u = b^2\frac{\mathrm{e}^{2at}-1}{2a} := \sigma^2(t)$$

and, since the probability density of a Gaussian process is determined by its expectation and variance process, we have $X(t) \sim \mathcal{N}(\mathrm{e}^{at}x,\ \sigma^2(t))$. The transition density of X is thus given by

$$p_X(t,\ x,\ s,\ z) = \frac{1}{\sqrt{2\pi\sigma^2(s-t)}}\exp\left(-\frac{(\mathrm{e}^{a(s-t)}x - z)^2}{2\sigma^2(s-t)}\right). \tag{5.4}$$

Similarly, for the reverse process Y we have $Y(s) \sim \mathcal{N}(\mathrm{e}^{-a(s-t)}y,\ \mathrm{e}^{-2a(s-t)}\sigma^2(s-t))$, and so

$$p_Y(t,\ y,\ s,\ z) = \frac{1}{\sqrt{2\pi\mathrm{e}^{-2a(s-t)}\sigma^2(s-t)}}\exp\left(-\frac{(\mathrm{e}^{-a(s-t)}y - z)^2}{2\mathrm{e}^{-2a(s-t)}\sigma^2(s-t)}\right)$$

is the transition density of Y.

We now consider the forward–reverse estimator (4.3) for the transition density (5.4), where we take $t = 0$ and $0 < t^* < T$. For the sake of simplicity, we do not deal with variance reduction, that is to say, we take $h \equiv 0$ and $F \equiv 0$. It follows that

$$p_X(0,\ x,\ T,\ y) \approx \xi_{N,M} := \frac{\mathrm{e}^{-a(T-t^*)}}{MN\delta}\sum_{m=1}^{M}\sum_{n=1}^{N} K_{nm}, \tag{5.5}$$

where

$$K_{nm} := K\left(\delta^{-1}\mathrm{e}^{at^*}\left(x + b\int_0^{t^*}\mathrm{e}^{-au}\,\mathrm{d}W^{(n)}(u)\right)\right.$$
$$\left. - \delta^{-1}\mathrm{e}^{-a(T-t^*)}\left(y + b\int_{t^*}^{T}\mathrm{e}^{a(u-t^*)}\,\mathrm{d}\tilde{W}^{(m)}(u)\right)\right)$$
$$= K(\delta^{-1}(\mathrm{e}^{at^*}x - \mathrm{e}^{-a(T-t^*)}y + \sigma(t^*)U^{(n)} - \mathrm{e}^{-a(T-t^*)}\sigma(T-t^*)V^{(m)})), \qquad (5.6)$$

with $U^{(n)}$ and $V^{(m)}$ being independent standard normally random variables. Note that, in general, δ in (5.5) and (5.6) may be chosen dependent on both N and M, so $\delta = \delta_{N,M}$ in fact.

By choosing the Gaussian kernel

$$K(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2), \qquad (5.7)$$

it is possible to derive explicit expressions for the first and second moment of $\xi_{N,M}$ in (5.5). In particular, for the expected value we have

$$\mathrm{E}\,\xi_{N,M} = \frac{1}{\sqrt{2\pi(\delta^2\mathrm{e}^{2a(T-t^*)} + \sigma^2(T))}}\exp\left(-\frac{(\mathrm{e}^{aT}x - y)^2}{2(\delta^2\mathrm{e}^{2a(T-t^*)} + \sigma^2(T))}\right) \qquad (5.8)$$

and for the variance it follows that

$$\mathrm{var}(\xi_{N,M}) = \frac{-N - M + 1}{2\pi MN(B + \sigma^2(T))}\exp\left(-\frac{A}{B + \sigma^2(T)}\right)$$
$$+ \frac{M-1}{2\pi MN\sqrt{B + \sigma^2(T-t^*)}\sqrt{B + 2\sigma^2(T) - \sigma^2(T-t^*)}}\exp\left(-\frac{A}{B + 2\sigma^2(T) - \sigma^2(T-t^*)}\right)$$
$$+ \frac{N-1}{2\pi MN\sqrt{B + \sigma^2(T) - \sigma^2(T-t^*)}\sqrt{B + \sigma^2(T) + \sigma^2(T-t^*)}}$$
$$\times\exp\left(-\frac{A}{B + \sigma^2(T) + \sigma^2(T-t^*)}\right) + \frac{\mathrm{e}^{-a(T-t^*)}}{2\pi MN\delta\sqrt{B + 2\sigma^2(T)}}\exp\left(-\frac{A}{B + 2\sigma^2(T)}\right). \qquad (5.9)$$

with the abbreviations $A := (\mathrm{e}^{aT}x - y)^2$, $B := \delta^2\mathrm{e}^{2a(T-t^*)}$. Since in Sections 6 the forward–reverse kernel estimator will be analysed quite generally, here we confine ourselves to a brief sketch of the derivation of (5.8) and (5.9). It is convenient to use the following standard lemma, which we state without proof.

Lemma 5.1. *Let U be a standard normal random variable and let the kernel K be given by* (5.7). *Then*

$$\mathrm{E}\, K(p+qU) = \frac{\exp(-p^2/(2+2q^2))}{\sqrt{2\pi(1+q^2)}}.$$

In (5.5) the K_{nm} are identically distributed and so (5.8) follows straightforwardly by application of Lemma 5.1. The variance expression can be derived as follows. We consider the second moment

$$\mathrm{E}\, \xi_{N,M}^2 = \frac{\mathrm{e}^{-2a(T-t^*)}}{M^2N^2\delta^2} \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{m'=1}^{M}\sum_{n'=1}^{N} \mathrm{E}\, K_{nm}K_{n'm'} \tag{5.10}$$

and split the sum into four parts: $n \neq n'$ and $m \neq m'$; $n = n'$ and $m \neq m'$; $n \neq n'$ and $m = m'$; $n = n'$ and $m = m'$. Then, to each part we apply Lemma 5.1 with appropriate substitutes for p and q. After collecting the results, (5.9) follows by $\mathrm{var}\,(\xi_{N,M}) = \mathrm{E}\, \xi_{N,M}^2 - (\mathrm{E}\, \xi_{N,M})^2$.

Clearly, as in Remark 4.1, substituting $t^* = T$ and $t^* = 0$ in (5.5) yields the pure forward estimator and pure reverse estimator, respectively. In this example the forward estimator is given by

$$\xi_N := \frac{1}{N\delta}\sum_{n=1}^{N} K_n := \frac{1}{N\delta}\sum_{n=1}^{N} K((\mathrm{e}^{aT}x - y + \sigma(T)U^{(n)})\delta^{-1})$$

and a similar expression holds for the reverse estimator. The bias and variance of these estimators may be derived analogously, but also follow from (5.8) and (5.9) by setting $t^* = T$ or $t^* = 0$.

We now compare the bias and variance of the forward–reverse estimator with the pure forward estimator. By (5.8) we have for the forward–reverse estimator, that is, (5.5) with $0 < t^* < T$,

$$\mathrm{E}\, \xi_{N,M} = \frac{\exp(-(\mathrm{e}^{aT}x - y)^2/2\sigma^2(T))}{\sqrt{2\pi\sigma^2(T)}}(1 + c_0\delta^2 + \mathcal{O}(\delta^3)) = p_X(0, x, T, y)(1 + \mathcal{O}(\delta^2)), \tag{5.11}$$

where c_0 is a constant not equal to zero. Hence, for a kernel given by (5.7) the bias is of order $\mathcal{O}(\delta^2)$. Obviously, the same is true for the forward estimator.

For the variance of the forward estimator we have

$$\mathrm{var}\,(\xi_N) = \frac{1}{2\pi N}\frac{\exp\,(-(\mathrm{e}^{aT}x - y)^2/(\delta^2 + 2\sigma^2(T)))}{\delta\sqrt{\delta^2 + 2\sigma^2(T)}} - \frac{1}{2\pi N}\frac{\exp(-(\mathrm{e}^{aT}x - y)^2/(\delta^2 + \sigma^2(T)))}{\delta^2 + \sigma^2(T)}, \tag{5.12}$$

which follows by substituting $t^* = T$ in (5.9) so that M drops out. Comparison of (5.9) with (5.12) then leads to the following interesting conclusion.

Conclusions 5.1. We consider the case $M = N$ and denote the forward–reverse estimator for

$p_X(0, x, T, y)$ by ξ_N as well. The width δ will thus be chosen in relation to N, hence $\delta = \delta_N$. We observe that

$$\mathrm{E}(\xi_N - p_X(0, x, T, y))^2 = \mathrm{var}(\xi_N) + (\mathrm{E}\,\xi_N - p_X(0, x, T, y))^2, \tag{5.13}$$

where $\varepsilon_N := \sqrt{\mathrm{E}(\xi_N - p_X(0, x, T, y))^2}$ is usually referred to as the accuracy of the estimation. From (5.11), (5.12) and (5.13) it is clear that for the forward estimator $\varepsilon_N \downarrow 0$ when $N \to \infty$, if and only if $\delta_N \to 0$ and $N\delta_N \to \infty$. By (5.11) and (5.12) we have for the forward estimator

$$\varepsilon_N^2 = \left(\frac{c_1}{N\delta_N} + c_2\delta_N^4\right)(1 + o(1)), \quad N\delta_N \to \infty \text{ and } \delta_N \downarrow 0, \tag{5.14}$$

for some positive constants c_1, c_2. It thus follows that the best achievable accuracy rate for the forward estimator is $\varepsilon_N \sim N^{-2/5}$, which is attained by taking $\delta_N \sim N^{-1/5}$.

We next consider the forward–reverse estimator which is obtained for $0 < t^* < T$. From (5.11), (5.9), and (5.13) it follows by similar arguments that

$$\varepsilon_N^2 = \left(\frac{d_1}{N} + \frac{d_2}{N^2\delta_N} + d_3\delta_N^4\right)(1 + o(1)), \quad N^2\delta_N \to \infty \text{ and } \delta_N \downarrow 0, \tag{5.15}$$

for some positive constants d_1, d_2 and d_3. So from (5.15) we conclude that by using the forward–reverse estimator the accuracy rate is improved to $\varepsilon_N \sim N^{-1/2}$ and this rate may be achieved by $\delta_N \sim N^{-p}$ for any $p \in [\frac{1}{4}, 1]$.

Remark 5.1. It is easy to check that for the reverse estimator we have the same accuracy (5.14) and so the same conclusions apply.

6. Accuracy analysis of the forward–reverse kernel estimator in general

In this section we study the properties of the kernel estimator (4.3) for the transition density $p = p(t, x, T, y)$ in general. However, here and in Section 7 we will disregard the discretization bias caused by numerical integration of SDEs and will only concentrate on the loss due to the particular structure of the new estimators. We thus assume in this section and the next that all random variables involved are due to exact solutions of the respective SDEs.

Let $r(u)$ be the density of the random variable $X_{t,x}(t^*)$, that is, $r(u) = p(t, x, t^*, u)$. Similarly, let $q(u)$ be the density of $Y_{t^*,y}(T)$ and further denote by $\mu(u)$ the conditional mean of $\mathcal{Y}_{t^*,y}(T)$ given $Y_{t^*,y}(T) = u$. By the following lemma we may reformulate the representation for p in (4.2) and $J(f)$ in (4.4).

Lemma 6.1.

$$p = \int r(u)\mu(u)q(u)\,\mathrm{d}u, \tag{6.1}$$

$$J(f) = \iint f(u, v)r(u)q(v)\mu(v)\,\mathrm{d}u\,\mathrm{d}v. \tag{6.2}$$

Proof. Equality (6.1) follows from (4.2) by

$$p = \mathrm{E}\, r\big(Y_{t^*,y}(T)\big)\mathcal{Y}_{t^*,y}(T) = \mathrm{E}\left[r\big(Y_{t^*,y}(T)\big)\mathrm{E}\left(\mathcal{Y}_{t^*,y}(T)|Y_{t^*,y}(T)\right)\right]$$

$$= \mathrm{E}\, r\big(Y_{t^*,y}(T)\big)\mu(Y_{t^*,y}(T)) = \int r(u)\mu(u)q(u)\,\mathrm{d}u, \tag{6.3}$$

and (6.2) follows from (4.4) in a similar way. □

For a kernel function $K(z)$ in $\mathbb{R}^d$ and a bandwidth δ, we put $f(u, v) = f_{K,\delta}(u, v) := \delta^{-d}K((u - v)/\delta)$ and thus have, by Lemma 6.1,

$$J(f_{K,\delta}) = \iint \delta^{-d}K\Big(\frac{u - v}{\delta}\Big)r(u)q(v)\mu(v)\,\mathrm{d}u\,\mathrm{d}v,$$

which formally converges to the target density p in (6.1) as $\delta \downarrow 0$. Following Section 4, this leads to the Monte Carlo kernel estimator

$$\hat{p} = \frac{1}{\delta^d MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\mathcal{Y}_m K\left(\frac{X_n - Y_m}{\delta}\right) = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}Z_{nm} \tag{6.4}$$

with

$$Z_{nm} := \delta^{-d}\mathcal{Y}_m K\left(\frac{X_n - Y_m}{\delta}\right),$$

where $X_n := X_{t,x}^{(n)}(t^*) \in \mathbb{R}^d$, $n = 1, \ldots, N$, may be regarded as an independent and identically distributed (i.i.d.) sample from the distribution with density r, the sequence $Y_m = Y_{t^*,y}^{(m)}(T) \in \mathbb{R}^d$, $m = 1, \ldots, M$, as an i.i.d. sample from the distribution with density q, and the weights $\mathcal{Y}_m = \mathcal{Y}_{t^*,y}^{(m)}(T)$, $m = 1, \ldots, M$, may be seen as independent samples from a distribution conditional on Y_m, with conditional mean $\mu(y)$ given $Y_m = u$. We derive some properties of this estimator below.

Lemma 6.2. *We have*

$$\mathrm{E}\,\hat{p} = p_\delta := \iint r(u + \delta v)q(u)\mu(u)K(v)\,\mathrm{d}u\,\mathrm{d}v = \int r_\delta(u)\lambda(u)\,\mathrm{d}u,$$

with

$$\lambda(u) := q(u)\mu(u)$$

and

$$r_\delta(u) := \delta^{-d}\int r(v)K\big(\delta^{-1}(v-u)\big)\,\mathrm{d}v = \int r(u+\delta v)K(v)\,\mathrm{d}v.$$

Moreover, if the kernel K satisfies $\int K(u)\,\mathrm{d}u = 1$, $K(u) \geqslant 0$, $K(u) = K(-u)$ for all $u \in \mathbb{R}^d$, and $K(u) = 0$ for $|u| > 1$, then the bias $|p - \mathrm{E}\,\hat{p}|$ satisfies

$$|p - \mathrm{E}\,\hat{p}| = |p - p_\delta| \leqslant C_K \|r''\|\delta^2 \tag{6.5}$$

with $C_K = \frac{1}{2}\int |v|^2 K(v)\,\mathrm{d}v \cdot \int \lambda(u)\,\mathrm{d}u$ and $\|r''\| = \sup_v \|r''(v)\|$, where $\|r''(v)\|$ is the Euclidean norm of the matrix $r''(v) = \{\partial^2 r/\partial v^i \partial v^j\}$.

Proof. Since all Z_{nm} are i.i.d., by (4.4) we have $\mathrm{E}\,\hat{p} = J(f_{K,\delta}) = \mathrm{E}\,Z_{nm}$ for every $n = 1, \ldots, N$ and $m = 1, \ldots, M$. Hence, by Lemma 6.1,

$$\begin{aligned}\mathrm{E}\,Z_{nm} &= \delta^{-d}\iint r(u)q(v)\mu(v)K\big(\delta^{-1}(u-v)\big)\,\mathrm{d}u\,\mathrm{d}v \\ &= \iint r(u+\delta v)q(u)\mu(u)K(v)\,\mathrm{d}u\,\mathrm{d}v = p_\delta.\end{aligned}$$

For the second assertion it is sufficient to note that the properties $\int K(v)\,\mathrm{d}v = 1$, $\int K(v)\,v\,\mathrm{d}v = 0$, and $K(v) = 0$ for $|v| > 1$ imply

$$\begin{aligned}r_\delta(u) - r(u) &= \int r(u+\delta v)K(v)\,\mathrm{d}v - r(u) = \int \big[r(u+\delta v) - r(u) - \delta v^{\mathrm{T}} r'(u)\big]K(v)\,\mathrm{d}v \\ &= \int \frac{1}{2}\delta^2 v^{\mathrm{T}} r''(u+\theta(v)\delta v)v\,K(v)\,\mathrm{d}v \\ &\leqslant \frac{1}{2}\delta^2\|r''\|\int |v|^2 K(v)\,\mathrm{d}v,\end{aligned}$$

where $|\theta(v)| \leqslant 1$, and so

$$|p_\delta - p| \leqslant \int |r_\delta(u) - r(u)|\lambda(u)\,\mathrm{d}u \leqslant C_K\delta^2\|r''\|\int \lambda(u)\,\mathrm{d}u.$$

□

Remark 6.1. The order of the bias $|p_{\delta-p}|$ can be improved by using higher-order kernels K. We say that K is of order β if $\int u_1^{j_1} \ldots u_d^{j_d} K(u)\,\mathrm{d}u = 0$ for all non-negative integers $j_1, \ldots, j_d$ satisfying $0 < j_1 + \ldots + j_d \leqslant \beta$. Similarly to the proof of Lemma 6.2, one can show that the application of a kernel K of order β satisfying $\int K(u)\,\mathrm{d}u = 1$, $K(u) = 0$ for $|u| \geqslant 1$ leads to a bias with $|p_\delta - p| \leqslant C\delta^{\beta+1}$, where C is a constant depending on r, q and K.

Concerning the variance $\mathrm{var}\,\hat{p} = \mathrm{E}(\hat{p} - \mathrm{E}\hat{p})^2$ of the estimator (6.4) we obtain the next result.

Lemma 6.3. *It holds*

$$\operatorname{var}\hat{p} = \frac{1}{NM}\delta^{-d}B_\delta + \frac{M-1}{NM}\int r(u)\lambda_\delta^2(u)\,\mathrm{d}u + \frac{N-1}{NM}\int r_\delta^2(u)\mu_2(u)q(u)\,\mathrm{d}u - \frac{N+M-1}{NM}\,p_\delta^2, \tag{6.6}$$

where

$$B_\delta = \int r_{\delta,2}(u)\mu_2(u)q(u)\,\mathrm{d}u$$

with

$$\lambda_\delta(u) = \delta^{-d}\int \lambda(v)K\big(\delta^{-1}(v-u)\big)\mathrm{d}v = \int \lambda(u+\delta v)K(v)\,\mathrm{d}v,$$

$$r_{\delta,2}(u) = \delta^{-d}\int r(v)K^2\big(\delta^{-1}(v-u)\big)\mathrm{d}v = \int r(u+\delta v)K^2(v)\,\mathrm{d}v,$$

$$\mu_2(v) = \mathrm{E}\,(\mathcal{Y}_1^2|\,Y_1 = v).$$

Proof. Since Z_{nm} and $Z_{n'm'}$ are independent if both $n \neq n'$ and $m \neq m'$, it follows that

$$M^2N^2\operatorname{var}\hat{p} = \mathrm{E}\left(\sum_{n=1}^{N}\sum_{m=1}^{M}(Z_{nm}-p_\delta)\right)^2 \tag{6.7}$$

$$= \sum_{n=1}^{N}\sum_{m=1}^{M}\mathrm{E}\,(Z_{nm}-p_\delta)^2 + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{m'\neq m}(\mathrm{E}\,Z_{nm}Z_{nm'} - p_\delta^2)$$

$$+ \sum_{n=1}^{N}\sum_{n'\neq n}\sum_{m=1}^{M}(\mathrm{E}\,Z_{nm}Z_{n'm} - p_\delta^2).$$

Note that for $m \neq m'$ we have

$$\mathrm{E}\,Z_{nm}Z_{nm'} = \delta^{-2d}\iiint K\big(\delta^{-1}(u-v)\big)K\big(\delta^{-1}(u-v')\big)r(u)\lambda(v)\lambda(v')\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}v'$$

$$= \delta^{-d}\iint K\big(\delta^{-1}(u-v)\big)r(u)\lambda_\delta(u)\lambda(v)\,\mathrm{d}u\,\mathrm{d}v$$

$$= \int r(u)\lambda_\delta^2(u)\,\mathrm{d}u$$

and, similarly, for $n \neq n'$ it follows that

$$\mathrm{E}\,Z_{nm}Z_{n'm} = \int r_\delta^2(u)\mu_2(u)q(u)\,\mathrm{d}u.$$

Further,

$$\begin{aligned}
\mathrm{E}\, Z_{nm}^2 &= \delta^{-2d}\mathrm{E}\, \mathcal{Y}_m^2 K^2\big(\delta^{-1}(X_n - Y_m)\big) \\
&= \delta^{-2d}\mathrm{E}\, \big(K^2\big(\delta^{-1}(X_n - Y_m)\big)\mathrm{E}\, \big(\mathcal{Y}_m^2 | Y_m\big)\big) \\
&= \delta^{-2d}\iint K^2\big(\delta^{-1}(u - v)\big)\, r(u) q(v) \mu_2(v)\, \mathrm{d}u\, \mathrm{d}v \\
&= \delta^{-d}\int \mu_2(v) q(v) r_{\delta,2}(v)\, \mathrm{d}v
\end{aligned}$$

and so we obtain

$$\operatorname{var} \hat{p} = \frac{\delta^{-d} B_\delta - p_\delta^2}{NM} + \frac{M-1}{NM}\left(\int r(u)\lambda_\delta^2(u)\, \mathrm{d}u - p_\delta^2\right) + \frac{N-1}{NM}\left(\int r_\delta^2(u)\mu_2(u)\, q(u)\, \mathrm{d}u - p_\delta^2\right),$$

from which the assertion follows. □

Let us define

$$B = \int K^2(u)\, \mathrm{d}u \cdot \int r(u)\mu_2(u)\, q(u)\, \mathrm{d}u. \tag{6.8}$$

By the Taylor expansion

$$r(u + \delta v) = r(u) + \delta v^{\mathrm{T}} r'(u) + \frac{1}{2}\delta^2 v^{\mathrm{T}} r''(u + \theta(v)\delta v) v,$$

one can show in a way similar to the proof of Lemma 6.1 that

$$|B_\delta - B| = O(\delta^2), \qquad \delta \downarrow 0.$$

In the same way, we obtain

$$\left|\int r(u)\lambda_\delta^2(u)\, \mathrm{d}u - \int r(u)\lambda^2(u)\, \mathrm{d}u\right| = O(\delta^2), \qquad \delta \downarrow 0,$$

$$\left|\int r_\delta^2(u)\mu_2(u)\, q(u)\, \mathrm{d}u - \int r^2(u)\mu_2(u)\, q(u)\, \mathrm{d}u\right| = O(\delta^2), \qquad \delta \downarrow 0.$$

Further, introduce the constant D given by

$$D := \int r(u)\lambda^2(u)\, \mathrm{d}u + \int r^2(u)\mu_2(u)\, q(u)\, \mathrm{d}u - 2p^2. \tag{6.9}$$

Then, from Lemmas 6.1 and 6.3 the next lemma follows.

Lemma 6.4. *For $N = M$ we have*

$$\left|\operatorname{var} \hat{p} - \frac{D}{N} - \frac{\delta^{-d} B}{N^2}\right| \leqslant C\left(\frac{\delta^{-d+2}}{N^2} + \frac{\delta^2}{N} + \frac{1}{N^2}\right). \tag{6.10}$$

In particular, if $\delta =: \delta_N$ *depends on* N *such that* $\delta_N^{-d} N^{-1} = o(1)$ *and* $\delta_N = o(1)$ *as* $N \to \infty$, *then*

$$\left| \operatorname{var} \hat{p} - \frac{D}{N} \right| = \frac{o(1)}{N}, \qquad N \to \infty.$$

Now, by combining Lemmas 6.2 and 6.4, we have the following theorem.

Theorem 6.1. *Let* $N = M$ *and* $\delta = \delta_N$ *depend on* N.

(a) *If* $d < 4$ *and* δ_N *is such that*

$$\frac{1}{N\delta_N^d} = o(1) \quad \textit{and} \quad \delta_N^4 N = o(1), \qquad N \to \infty,$$

then the estimate $\hat{p}$ *(see* (4.3) *or* (6.4)*) of the transition density* $p = p(t, x, T, y)$ *satisfies*

$$\mathrm{E}\,(\hat{p} - p)^2 = (p_\delta - p)^2 + \operatorname{var} \hat{p} = \frac{D}{N} + \frac{o(1)}{N}, \qquad N \to \infty. \tag{6.11}$$

Hence, a root-N accuracy rate is achieved (we recall that $\sqrt{\mathrm{E}\,(\hat{p} - p)^2}$ *is the accuracy of the estimator). In this case the variance is of order* N^{-1} *and the squared bias is* $o(N^{-1})$.

(b) *If* $d = 4$ *and* $\delta_N = CN^{-1/4}$, *where* C *is a positive constant, then the accuracy rate is again* $N^{-1/2}$ *but now both the squared bias and the variance are of order* N^{-1}.

(c) *If* $d > 4$ *and* $\delta_N = CN^{-2/(4+d)}$, *then the accuracy rate is* $N^{-4/(4+d)}$ *and both the squared bias and the variance are of the same order* $N^{-8/(4+d)}$.

Proof. Clearly, (6.5) and (6.10) imply (6.11). The conditions $\delta_N^{-d} N^{-1} = o(1)$ and $N\delta_N^4 = o(1)$ can be fulfilled simultaneously only when $d < 4$. In this case one may take, for instance, $\delta_N = N^{-1/d} \log^{1/d} N$ yielding $\delta_N^{-d} N^{-1} = 1/\log N = o(1)$ and $N\delta_N^4 = N^{1-4/d} \log^{4/d} N = o(1)$. By (6.5) the squared bias is then of order $\mathcal{O}(\delta_N^4) = \mathcal{O}(N^{-4/d} \log^{4/d} N) = o(N^{-1})$ for $d < 4$. The statements for $d = 4$ and $d > 4$ follow in a similar way. □

Remark 6.2. We conclude that, by combining forward and reverse diffusion, it is possible to achieve an estimation accuracy of rate $N^{-1/2}$ for $d \leqslant 4$. Moreover, for $d > 4$ an accuracy rate of root-N may also be achieved by applying a higher-order kernel K.

In Section 9 we will see that with the proposed choice of the bandwidth $\delta N = N^{-1/d} \log^{1/d} N$ for $d \leqslant 3$ and $\delta_N = N^{-2/(4+d)}$ for $d \geqslant 4$, the kernel estimator $\hat{p}$ can be computed at a cost of order $N \log N$ operations.

Remark 6.3. For the pure forward estimator (1.6) and pure reverse estimator (1.6) it is not difficult to show that

$$\varepsilon_{\mathrm{N}}^2 := \mathrm{E}\,(\hat{p} - p)^2 = \left(\frac{c_1}{N\delta_N^d} + c_2\delta_N^4\right)(1 + o(1)), \qquad \delta_N \downarrow 0 \text{ and } N\delta_N^d \to \infty, \tag{6.12}$$

where c_1 and c_2 are positive constants. So the best achievable accuracy rate for the forward estimator is $\varepsilon_N = \mathcal{O}(N^{-2/(4+d)})$, which is obtained by a bandwidth choice $\delta_N = N^{-1/(4+d)}$. Clearly, this rate is lower than the accuracy rate of the forward–reverse estimator which is basically root-N.

Remark 6.4. In applications it is important to choose the intermediate time t^* properly. In this respect we note that D in (6.9) only depends on the choice of t^* and, in particular, it is not difficult to show that $D \to \infty$ as $t^* \downarrow t$ or $t^* \uparrow T$. So, by Lemma 6.4, in the case $N = M$ and $d < 4$ we should select a t^* for which this constant is not too big. In practice, however, a suitable t^* is best found by just comparing for different choices the performance of the estimator for relatively small sample sizes. For $d \geqslant 4$ and $N = M$ the constant B in (6.8) is also involved but similar conclusions can be drawn.

7. The forward–reverse projection estimator

In this section we discuss statistical properties of the *projection* estimator $\hat{p}^{\mathrm{pr}}$ from (4.6) for the transition density $p(t, x, T, y)$. First we sketch the main idea.

Let $\{\varphi_\ell(x), \ell = 1, 2, \ldots\}$ be a total orthonormal system in the Hilbert space $L_2(\mathbb{R}^d)$. For example, in the case $d = 1$ one could take

$$\varphi_{l+1}(u) = \frac{1}{\sqrt{2^l l!}\sqrt[4]{\pi}} H_l(u)\mathrm{e}^{-u^2/2}, \qquad l = 0, 1, \ldots,$$

where $H_l(u)$ are the Hermite polynomials. In the d-dimensional case it is possible to construct a similar basis by using Hermite functions as well. Consider formally for $r(u) = p(t, x, t^*, u)$ (see Section 6) and $h(u) := p(t^*, u, T, y)$ the Fourier expansions

$$r(u) = \sum_{\ell=1}^{\infty} \alpha_\ell \varphi_\ell(u), \qquad h(u) = \sum_{\ell=1}^{\infty} \gamma_\ell \varphi_\ell(u),$$

with

$$\alpha_\ell := \int r(u)\varphi_\ell(u)\,\mathrm{d}u, \qquad \gamma_\ell := \int h(u)\varphi_\ell(u)\,\mathrm{d}u.$$

By (2.1), (3.1), and (3.5) it follows that

$$\alpha_\ell = \mathrm{E}\,\varphi_\ell(X_{t,x}(t^*)), \tag{7.1}$$

$$\gamma_\ell = \mathrm{E}\,\varphi_\ell(Y_{t^*,y}(T))\mathcal{Y}_{t^*,y}(T), \tag{7.2}$$

respectively. Since by the Chapman–Kolmogorov equation (4.1) the transition density $p = p(t, x, T, y)$ may be written as a scalar product $p = \int r(u)h(u)\,\mathrm{d}u$, we thus formally obtain

$$p = \sum_{\ell=1}^{\infty} \alpha_\ell \gamma_\ell. \tag{7.3}$$

Therefore, it is natural to consider the estimator

$$\hat{p}^{\mathrm{pr}} = \sum_{\ell=1}^{L} \hat{\alpha}_\ell \hat{\gamma}_\ell, \tag{7.4}$$

where L is a natural number and

$$\hat{\alpha}_\ell := \frac{1}{N} \sum_{n=1}^{N} \varphi_\ell(X_n), \qquad \hat{\gamma}_\ell := \frac{1}{M} \sum_{m=1}^{M} \varphi_\ell(Y_m) \mathcal{Y}_m \tag{7.5}$$

are estimators for the Fourier coefficients α_ℓ, γ_ℓ, respectively. For the definition of X_n, Y_m and $\mathcal{Y}_m$, see Section 6. Note that (7.4)–(7.5) coincide with the projection estimator introduced in (4.6).

We now study the accuracy of the projection estimator. In the subsequent analysis we assume that the originating diffusion coefficients a and σ in (1.1) are sufficiently good in the analytical sense such that, in particular, the functions $y' \to p(t, x, t^*, y')$ and $y' \to p(t^*, y', T, y)$ are square integrable. Hence, we assume that the Fourier expansions used in this section are valid in $L_2(\mathbb{R}^d)$. The notation introduced in Section 6 is retained below. We have the following lemma.

Lemma 7.1. *For every* $\ell \geqslant 1$,

$$\mathrm{E}\,\hat{\alpha}_\ell = \alpha_\ell = \int r(u) \varphi_\ell(u)\,\mathrm{d}u,$$

$$\mathrm{var}\,\hat{\alpha}_\ell = N^{-1}\,\mathrm{var}\,\varphi_\ell(X_1) = N^{-1}\left(\int \varphi_\ell^2(u) r(u)\,\mathrm{d}u - \alpha_\ell^2\right) =: N^{-1}\alpha_{\ell,2}.$$

Similarly,

$$\mathrm{E}\,\hat{\gamma}_\ell \;=\; \gamma_\ell = \int \varphi_\ell(u)\mu(u)q(u)\,\mathrm{d}u,$$

$$\mathrm{var}\,\hat{\gamma}_\ell \;=\; M^{-1}\,\mathrm{var}\,\mathcal{Y}_1\varphi_\ell(Y_1) = M^{-1}\left(\int \mu_2(u)\varphi_\ell^2(u)q(u)\,\mathrm{d}u - \gamma_\ell^2\right) =: M^{-1}\gamma_{\ell,2},$$

where $\mu_2(u) := \mathrm{E}\,(\mathcal{Y}_1^2 | Y_1 = u)$.

Proof. The first part is obvious and the second part follows by a conditioning argument similar to (6.3) in the proof of Lemma 6.1. □

Since the $\hat{\alpha}_\ell$ and the $\hat{\gamma}_\ell$ are independent, it follows by Lemma 7.1 that

$$\mathrm{E}\,\hat{p}^{\mathrm{pr}} = \mathrm{E}\sum_{\ell=1}^{L}\hat{\alpha}_\ell\hat{\gamma}_\ell = \sum_{\ell=1}^{L}\alpha_\ell\gamma_\ell.$$

So, by (7.3) and the Cauchy–Schwarz inequality, we obtain the next lemma for the bias $\mathrm{E}\,\hat{p}^{\mathrm{pr}} - p$ of the estimator $\hat{p}^{\mathrm{pr}}$.

Lemma 7.2.

$$(\mathrm{E}\,\hat{p}^{\mathrm{pr}} - p)^2 = \left(\sum_{\ell=L+1}^{\infty}\alpha_\ell\gamma_\ell\right)^2 \leqslant \sum_{\ell=L+1}^{\infty}\alpha_\ell^2\sum_{\ell=L+1}^{\infty}\gamma_\ell^2.$$

By the following result we may estimate the variance of $\hat{p}^{\mathrm{pr}}$. For convenience, we restrict ourselves to the case $N = M$.

Lemma 7.3. *Let $(L+1)^2 \leqslant N$ and the Fourier coefficients α_ℓ and γ_ℓ satisfy the conditions*

$$\sum_{\ell=1}^{\infty}|\alpha_\ell| \leqslant C_{1,\alpha}, \qquad \sum_{\ell=1}^{\infty}|\gamma_\ell| \leqslant C_{1,\gamma}, \tag{7.6}$$

$$\max_\ell \alpha_{\ell,2} \leqslant C_{2,\alpha}, \qquad \max_\ell \gamma_{\ell,2} \leqslant C_{2,\gamma}. \tag{7.7}$$

Then we have

$$N \operatorname{var} \hat{p}^{pr} \leqslant C,$$

with C depending on $C_{1,\alpha}$, $C_{2,\alpha}$ and $C_{1,\gamma}$, $C_{2,\gamma}$ only.

Proof. Let us write

$$\sum_{\ell=1}^{L}\hat{\alpha}_\ell\hat{\gamma}_\ell - \sum_{\ell=1}^{L}\alpha_\ell\gamma_\ell = \sum_{\ell=1}^{L}(\hat{\alpha}_\ell - \alpha_\ell)(\hat{\gamma}_\ell - \gamma_\ell) + \sum_{\ell=1}^{L}\alpha_\ell(\hat{\gamma}_\ell - \gamma_\ell) + \sum_{\ell=1}^{L}(\hat{\alpha}_\ell - \alpha_\ell)\gamma_\ell$$

$$=: I_1 + I_2 + I_3.$$

The Cauchy–Schwarz inequality implies that

$$\mathrm{E}\,(I_2)^2 = \mathrm{E}\left(\sum_{\ell=1}^{L}\alpha_\ell(\hat{\gamma}_\ell - \gamma_\ell)\right)^2 \leqslant \mathrm{E}\left(\sum_{\ell=1}^{L}|\alpha_\ell|\sum_{\ell=1}^{L}|\alpha_\ell|(\hat{\gamma}_\ell - \gamma_\ell)^2\right)$$

$$\leqslant C_{1,\alpha}\sum_{\ell=1}^{L}|\alpha_\ell|\mathrm{E}\,(\hat{\gamma}_\ell - \gamma_\ell)^2 \leqslant C_{1,\alpha}^2 C_{2,\gamma}N^{-1},$$

and similarly

$$\mathrm{E}\,(I_3)^2 = \mathrm{E}\left(\sum_{\ell=1}^{L}\gamma_\ell(\hat\alpha_\ell-\alpha_\ell)\right)^2 \leqslant C_{1,\gamma}^2 C_{2,\alpha} N^{-1}.$$

The Cauchy–Schwarz inequality and independence of the $\hat\alpha_\ell$ and the $\hat\gamma_\ell$ imply that

$$\mathrm{E}\,(I_1)^2 = \mathrm{E}\left(\sum_{\ell=1}^{L}(\hat\alpha_\ell-\alpha_\ell)(\hat\gamma_\ell-\gamma_\ell)\right)^2 \leqslant \mathrm{E}\sum_{\ell=1}^{L}(\hat\alpha_\ell-\alpha_\ell)^2\mathrm{E}\sum_{\ell=1}^{L}(\hat\gamma_\ell-\gamma_\ell)^2$$
$$\leqslant C_{2,\alpha}C_{2,\gamma}(L+1)^2N^{-2} \leqslant C_{2,\alpha}C_{2,\gamma}N^{-1}.$$

Hence,

$$\mathrm{var}\,\hat p^{\mathrm{pr}} = \mathrm{E}\,(I_1+I_2+I_3)^2 \leqslant \left(\sqrt{E(I_1)^2}+\sqrt{E(I_2)^2}+\sqrt{E(I_3)^2}\right)^2 \leqslant \frac{C}{N}$$

with $C := 3(C_{1,\alpha}^2C_{2,\gamma} + C_{1,\gamma}^2C_{2,\alpha} + C_{2,\alpha}C_{2,\gamma})$. □

Application of Lemmas 7.2 and 7.3 yields the following theorem.

Theorem 7.1. *Let the Fourier coefficients α_ℓ and γ_ℓ satisfy the condition*

$$\sum_{\ell=1}^{\infty}\alpha_\ell^2\ell^{2\beta/d} \leqslant C_\alpha^2, \qquad \sum_{\ell=1}^{\infty}\gamma_\ell^2\ell^{2\beta/d} \leqslant C_\gamma^2 \tag{7.8}$$

with $\beta > d/2$, and let condition (7.7) *hold. Let $L = L_N$ satisfy $L_N^2/N = o(1)$ and $NL_N^{-4\beta/d} = o(1)$, as $N \to \infty$. Then, for the accuracy of the estimator $\hat p^{\mathrm{pr}}$ with $N = M$, we have*

$$\mathrm{E}\,(\hat p^{\mathrm{pr}} - p)^2 \leqslant CN^{-1}.$$

Proof. Clearly,

$$\sum_{\ell=L+1}^{\infty}\alpha_\ell^2 \leqslant (L+1)^{-2\beta/d}\sum_{\ell=L+1}^{\infty}\alpha_\ell^2\ell^{2\beta/d} \leqslant C_\alpha^2L^{-2\beta/d}.$$

Similarly, $\sum_{\ell=L+1}^{\infty}\gamma_\ell^2 \leqslant C_\gamma^2L^{-2\beta/d}$ and so

$$N\left(\sum_{\ell=L+1}^{\infty}\alpha_\ell\gamma_\ell\right)^2 \leqslant C_\alpha^2C_\gamma^2NL^{-4\beta/d} = o(1).$$

Next,

$$\left(\sum_{\ell=1}^{L}|\alpha_\ell|\right)^2 \leqslant \sum_{\ell=1}^{L}\alpha_\ell^2\ell^{2\beta/d}\sum_{\ell=1}^{L}\ell^{-2\beta/d} \leqslant C_\alpha^2\sum_{\ell=1}^{L}\ell^{-2\beta/d} \leqslant C_\alpha^2C_\beta$$

with $C_\beta = \sum_{\ell=1}^{L}\ell^{-2\beta/d} < \infty$. Similarly

$$\left(\sum_{\ell=1}^{L}|\gamma_\ell|\right)^2 \leqslant C_\gamma^2 C_\beta,$$

and thus condition (7.6) holds with $C_{1,\alpha} = C_\alpha C_\beta^{1/2}$ and $C_{1,\gamma} = C_\gamma C_\beta^{1/2}$. Now the assertion follows from Lemma 7.3. □

Remark 7.1. In Theorem 7.1, β plays the role of a smoothness parameter. Indeed, for a functional basis such as the Hermite bases, condition (7.8) is fulfilled if the functions $x' \to p(t, x, t^*, x')$ and $x' \to p(t^*, x', T, y)$ have square-integrable derivatives up to order β. For $\beta = 2$, the conditions $L_N^2/N = o(1)$ and $NL_N^{-4\beta/d} = o(1)$ can be fulfilled simultaneously only if $d < 4$, so we then have a similar situation to that for the kernel estimator in Section 6. In general, if (7.8) holds for $\beta > d/2$, one may take $L_N = (N\log N)^{d/(4\beta)}$ in Theorem 7.1, thus yielding $L_N^2/N = N^{-1+d/(2\beta)}\log^{d/(2\beta)}N = o(1)$ and $NL_N^{-4\beta/d} = \log^{-1}N = o(1)$. However, with respect to sufficiently regular basis functions (such as Hermite basis functions) condition (7.8) is fulfilled for any $\beta > d/2$ when the densities $p(t, x, t^*, x')$ and $p(t^*, x', T, y)$ have square-integrable derivatives up to any order. So, according to Theorem 7.1, one could take $L_N = \mathcal{O}(N^\tau)$ for any $0 < \tau < 1/2$ to get the desired root-N consistency. If, moreover, the coefficients α_ℓ and γ_ℓ decrease exponentially fast so that $\sum_\ell \alpha_\ell e^{c\ell} < \infty$ and $\sum_\ell \gamma_\ell e^{c\ell} < \infty$ for some positive c (which corresponds to the case of analytical densities $p(t, x, t^*, x')$ and $p(t^*, x', T, y)$), then even $L_N = \mathcal{O}(\log N)$ Fourier coefficients provide a negligible estimation bias (see Pinsker 1980), thus leading to root-N consistency again. Generally it is clear that properly choosing L_N is essential for reducing the numerical complexity of the procedure (see Section 9).

Remark 7.2. The conditions of Theorem 7.1 are given in terms of the Fourier coefficients α_ℓ and γ_ℓ. We do not investigate in a rigorous way how these conditions can be transformed into conditions on the coefficients of the original diffusion model (1.1) and the chosen orthonormal basis. Note, however, that in the case of, for example, the Hermite basis, both (7.7) and (7.8) follow from standard regularity conditions. For instance, when the coefficients of (1.1) are smooth and bounded, their derivatives are smooth and bounded, and the matrix $\sigma(s, x)\sigma^{\mathrm{T}}(s, x)$ is of full rank for all s, x.

8. Estimation loss caused by numerical integration of SDEs

In this section we analyse the estimation loss of the kernel estimators due to application of the Euler scheme. Let $\overline{X} := \overline{X}_{t,x}(t^*, h)$ and $(\overline{Y}, \overline{\mathcal{Y}}) := (\overline{Y}_{t^*,y}(T, h), \overline{\mathcal{Y}}_{t^*,y}(T, h))$ be an approximation of $X_{t,x}(t^*)$ and $(Y_{t^*,y}(T), \mathcal{Y}_{t^*,y}(T))$, obtained by applying the Euler scheme to the systems (1.1) and (3.6), respectively. Let $\bar{r}(u)$ be the density of the random variable $\overline{X}$, so $\bar{r}(u) = \bar{p}_h(t, x, t^*, u)$. Further, let $\bar{q}(u)$ be the density of $\overline{Y}$ and denote by $\bar{\mu}(u)$ the conditional mean of $\overline{\mathcal{Y}}$ given $\overline{Y} = u$. Instead of (6.4) we now consider the estimator

$$\hat{\bar{p}} := \frac{1}{\delta^d MN} \sum_{n=1}^{N} \sum_{m=1}^{M} \overline{\mathcal{Y}}_m K\left(\frac{\overline{X}_n - \overline{Y}_m}{\delta}\right) = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} \overline{Z}_{nm}, \tag{8.1}$$

where

$$\overline{Z}_{nm} := \delta^{-d} \overline{\mathcal{Y}}_m K\left(\frac{\overline{X}_n - \overline{Y}_m}{\delta}\right),$$

with $\overline{X}_n$, $n = 1, \ldots, N$, and $(\overline{Y}_m, \overline{\mathcal{Y}}_m)$ $m = 1, \ldots, M$, being independent realizations of $\overline{X}$ and $(\overline{Y}, \overline{\mathcal{Y}})$, respectively. We thus have

$$\begin{aligned} \mathrm{E}\, \hat{\bar{p}} = \mathrm{E}\, \overline{Z}_{nm} &= \delta^{-d} \iint \bar{r}(u)\bar{q}(v)\bar{\mu}(v) K(\delta^{-1}(u - v))\, \mathrm{d}u\, \mathrm{d}v \\ &= \iint \bar{r}(u + \delta v)\bar{q}(u)\bar{\mu}(u) K(v)\, \mathrm{d}u\, \mathrm{d}v \\ &= \int \bar{r}_\delta(u)\bar{q}(u)\bar{\mu}(u)\, \mathrm{d}u, \end{aligned} \tag{8.2}$$

where

$$\bar{r}_\delta(u) := \int \bar{r}(u + \delta v) K(v)\, \mathrm{d}v.$$

From the result due to Bally and Talay (1996b) (see (1.4)) we obtain

$$|\bar{r}_\delta(u) - r_\delta(u)| \leqslant Kh, \tag{8.3}$$

uniform in u and δ for some positive constant K. Hence, for some $K_1 > 0$,

$$|\mathrm{E}\, \hat{\bar{p}} - \int r_\delta(u)\bar{q}(u)\bar{\mu}(u)\, \mathrm{d}u| \leqslant K_1 h. \tag{8.4}$$

uniform in δ. Further, we have

$$\int r_\delta(u)\bar{q}(u)\bar{\mu}(u)\, \mathrm{d}u = \mathrm{E}\, r_\delta(\overline{Y})\overline{\mathcal{Y}}. \tag{8.5}$$

It is not difficult to show that $r_\delta(u)$ has derivatives which are uniformly bounded with respect to δ. Therefore, since the Euler scheme has weak order 1, we have, for some $K_2 > 0$,

$$|\mathrm{E}\, r_\delta(\overline{Y})\overline{\mathcal{Y}} - \mathrm{E}\, \hat{p}| \leqslant K_2 h, \tag{8.6}$$

uniform in δ. Combining (8.4)–(8.6) yields

$$|\mathrm{E}\, \hat{\bar{p}} - \mathrm{E}\, \hat{p}| \leqslant K_3 h, \tag{8.7}$$

uniform in δ for some $K_3 > 0$, and then by Lemma 6.2 we obtain the following result.

Lemma 8.1. *The estimation loss* $|\mathrm{E}\, \hat{\bar{p}} - p|$ *satisfies*

$$|\mathrm{E}\, \hat{\bar{p}} - p| \leqslant K_4 \delta^2 + K_5 h,$$

for some positive constants K_4, K_5 independent of δ and h.

We now proceed with estimation of var $\hat{\bar{p}}$. For var $\hat{\bar{p}}$ we obtain an expression similar to (6.6) by replacing p_δ in (6.6) with $\bar{p}_\delta := \mathrm{E}\,\hat{\bar{p}}$ and throughout Lemma 6.3 the quantities r, r_δ, $r_{\delta,2}$, q, μ_2, λ, λ_δ, B_δ by their corresponding analogues $\bar{r}$, $\bar{r}_\delta$, $\bar{r}_{\delta,2}$, $\bar{q}$, $\bar{\mu}_2$, $\bar{\lambda}$, $\bar{\lambda}_\delta$, $\bar{B}_\delta$ defined with respect to the random variables $\overline{X}$ and $(\overline{Y}, \overline{\mathcal{Y}}$. Analogously to the proof of (8.7), it follows that, for some positive constants C, C_1,

$$|\bar{B}_\delta - B_\delta| \leqslant Ch \quad \text{and} \quad \left| \int \bar{r}_\delta^2(u)\bar{\mu}_2(u)\bar{q}(u)\,\mathrm{d}u - \int r_\delta^2(u)\mu_2(u)q(u)\,\mathrm{d}u \right| \leqslant C_1 h,$$

uniform in δ. From our boundedness assumptions in Section 1, it follows that $c(s, y)$ in (3.6) is bounded (see (3.4)). As a consequence, $\overline{\mathcal{Y}}_{t^*,y}(T)$ is bounded and so there exists a constant $C_2 > 0$ such that, for every h and u,

$$|\bar{\mu}(u)| = |\mathrm{E}\,(\overline{\mathcal{Y}}_{t^*,y}(T) | \overline{Y}_{t^*,y}(T) = u)| \leqslant C_2.$$

Therefore,

$$|\bar{\lambda}_\delta(u)| = \left| \int \bar{q}(u + \delta v)\bar{\mu}(u + \delta v)K(v)\,\mathrm{d}v \right| \leqslant C_3 \int \bar{q}(u + \delta v)K(v)\,\mathrm{d}v \tag{8.8}$$

for some $C_3 > 0$ and all u, h, δ.

By Bally and Talay (1996b) again, $\bar{q}(u) - q(u) = O(h)$ uniform in u; hence, $\bar{\lambda}_\delta(u)$ is uniformly bounded with respect to u, h and δ, and so $\int \bar{r}(u)\bar{\lambda}_\delta^2(u)\,\mathrm{d}u$ is uniformly bounded with respect to h and δ. Now, from Lemma 6.3 and the above arguments the following result is obvious.

Lemma 8.2. *There exist positive constants C_4 and C_5, not depending on h and δ, such that for $N = M$,*

$$\mathrm{var}\,\hat{\bar{p}} \leqslant \frac{C_4}{N^2\delta^d} + \frac{C_5}{N}. \tag{8.9}$$

It should be noted that Lemma 6.4 is more refined than Lemma 8.2 in the sense that it gives some kind of expansion of var $\hat{p}$. Nevertheless, it is clear that Lemmas 8.1 and 8.2 are sufficient to obtain the following main theorem.

Theorem 8.1. *for $M = N$ and positive constants D, D_1, D_2, D_3 we have*

$$\mathrm{E}\,(\hat{\bar{p}} - p)^2 \leqslant D\delta^4 + D_1 h^2 + \frac{D_2}{N^2\delta^d} + \frac{D_3}{N}. \tag{8.10}$$

Let us take $\delta = \delta_N$ as in Theorem 6.1. Then it is clear from Theorem 6.1 that for $d \leqslant 4$ and $h = O(N^{-1/2})$ the accuracy of the estimator $\hat{\bar{p}}$ is $O(N^{-1/2})$, and for $d > 4$ and $h = O(N^{-4/(4+d)})$ the accuracy of $\hat{\bar{p}}$ is $O(N^{-4/(4+d)})$. Hence by properly choosing h dependent on N the accuracy rates for $\hat{\bar{p}}$ and $\hat{p}$ coincide.

Remark 8.1. For the pure forward estimator (1.7) (and the pure reverse estimator corresponding to (4.7)) similar (but simpler) arguments give

$$\mathrm{E}\,(\hat{\hat{p}} - p)^2 \leqslant D_4 h^2 + D_5 \delta^4 + \frac{D_6}{N\delta^d}, \tag{8.11}$$

for positive constants D_4, D_5, D_6. For comparison see also Remark 6.3.

Remark 8.2. The assertions of this section are derived only for the Euler method in the strong sense since we essentially use the results of Bally and Talay (1996b). Most likely they remain true in the context of methods of numerical integration in a weak sense. However, this requires additional investigation.

Remark 8.3. Without proof we note that for the projection estimators similar conclusions can be drawn with respect to the estimation loss due to application of the Euler scheme.

9. Implementation of the forward–reverse estimators

In the previous sections we have shown that both the forward–reverse kernel and projection estimator have superior convergence properties compared with the classical Parzen–Rosenblatt estimator. However, while the implementation of the classical estimator is rather straightforward, one has to be more careful when implementing the forward–reverse estimation algorithms. This especially concerns the evaluation of the double sum in (4.3) for the kernel estimation. Indeed, straightforward computation would require MN kernel evaluations, which would be prohibitive, for example, when $M = N = 10^5$. Fortunately, by using kernels with small support, in some sense, we can get around this difficulty as outlined below.

9.1. Implementation of the kernel estimator and its numerical complexity

We assume here that the kernel $K(x)$ used in (4.3) has a small support contained in $|x|_{\max} \leqslant \alpha/2$ for some $\alpha > 0$, where $|x|_{\max} := \max_{1 \leqslant i \leqslant d} |x^i|$. This assumption is easily satisfied in practice. For instance, for the Gaussian kernel, $K(x) = (2\pi)^{-\mathrm{d}/2} \exp(-|x|^2/2)$, which, strictly speaking, has unbounded support, in practice $K(x)$ is negligible if for some i, $1 \leqslant i \leqslant d$, $|x_i| > 6$ and so we could take for this kernel $\alpha = 12$. Then, due to the small support of K, the following Monte Carlo algorithm for the forward–reverse kernel estimator is possible. For simplicity, we take $t = 0$, $t^* = T/2$ and assume $N = M$. For both forward and reverse trajectory simulation we use the Euler scheme with time discretization step $h = T/(2L)$, with $2L$ being the total number of steps between 0 and T.

- Simulate N trajectories on the interval $[0, t^*]$, with end points $\{X^{(n)}(t^*) : n = 1, \ldots, N\}$, at a cost of $\mathcal{O}(NLd)$ elementary computations.

- Simulate N reverse trajectories on the interval $[t^*, T]$, with end-points $\{(Y^{(m)}(T), \mathcal{Y}^{(m)}(T)) : m = 1, \ldots, N\}$ at a cost of $\mathcal{O}(NLd)$ elementary computations.
- Search, for each m, the subsample

$$\{X^{(n_k)}(t^*) : k = 1, \ldots, l_m\} := \{X^{(n)}(t^*) : n = 1, \ldots, N\} \cap \{x : |x - Y^{(m)}(T)|_{\max} \leqslant \alpha\delta_N\}.$$

  The size l_m of this intersection is, on average, approximately $N\delta_N^d \times \{\text{density of } X(t^*) \text{ at } Y^{(m)}(T)\}$. This search procedure can be done at a cost of order $\mathcal{O}(N \log N)$; see, for instance, Greengard and Strain (1991) where this is proved in the context of the Gauss transform.

- Finally, evaluate (4.3) by

$$\frac{1}{N^2\delta_N^d} \sum_{m=1}^{N} \sum_{k=1}^{l_m} K(\delta_N^{-1}(X^{(n_k)}(t^*) - Y^{(m)}(T)))\mathcal{Y}^{(m)}(T),$$

  at an estimated cost of $\mathcal{O}(N^2\delta_N^d)$.

For the study of complexity we use the results in Section 6. We distinguish between $d < 4$ and $d \geqslant 4$. For $1 \leqslant d < 4$ we achieve root-N accuracy by choosing $\delta_N = (N/\log N)^{-1/d}$. In practice, the number of discretization steps $2L$ (typically 100–1000) is much smaller than the Monte Carlo number N, which is typically 10^5–10^6. Therefore, as we see from the above algorithm, with $\delta_N = (N/\log N)^{-1/d}$ simulation of the forward–reverse estimator incurs a total cost of $\mathcal{O}(N \log N)$. Hence, the aggregated costs for achieving $\varepsilon_N \sim 1/\sqrt{N}$ amount to $\mathcal{O}(N \log N)$ which comes down to a complexity $C_\varepsilon^{\text{kern}} \sim |\log \varepsilon|/\varepsilon^2$. For $d \geqslant 4$ we achieve an accuracy rate $\varepsilon_N \sim N^{-4/(4+d)}$ by taking $\delta_N = N^{-2/(4+d)}$, again at a cost of $\mathcal{O}(N \log N)$. So the complexity $C_\varepsilon^{\text{kern}}$ is of order $\mathcal{O}(|\log \varepsilon|/\varepsilon^{(4+d)/4})$. For comparison we now consider the classical estimator. It is well known (see also Remark 6.3) that for N trajectories the optimal bandwidth choice is $\delta_N \sim N^{-1/(4+d)}$, which yields an accuracy of $\varepsilon_N \sim N^{-2/(4+d)}$. The costs of the classical estimator amounts to $\mathcal{O}(N)$ and thus its complexity $C_\varepsilon^{\text{class}}$ is of order $\mathcal{O}(1/\varepsilon^{(4+d)/2})$. By comparing the complexities C_ε and $C_\varepsilon^{\text{class}}$ it is clear that the forward–reverse kernel estimator is superior to the classical Parzen–Rosenblatt kernel estimator for any d.

9.2. Complexity of the projection estimator

From its construction in Section 7 it is clear that the evaluation of the projection estimator (4.6) incurs a cost of order $\mathcal{O}(L_N N)$ elementary computations. Just as for the kernel estimator, we now consider the complexity of the projection estimator. In Remark 7.1 we saw that if condition (7.8) is fulfilled for a smoothness β with $\beta > d/2$, we may choose $L_N = (N \log N)^{d/(4\beta)}$, which yields a complexity $C^{\text{proj}}(\varepsilon)$ of order $\mathcal{O}(\log^{d/(4\beta)}|\varepsilon|/\varepsilon^{2+d/(2\beta)})$. If, moreover, the Fourier coefficients α_ℓ and γ_ℓ decrease exponentially, then (see Remark 7.1) we achieve root-N accuracy by taking $L_N = \log N$ and so we obtain a complexity of

order $C^{\text{proj}}(\varepsilon) = |\log \varepsilon|/\varepsilon^2$ for any d. Obviously, compared to the classical estimator, the projection estimator has in any case a better order of complexity.

Remark 9.1. For transparency, the complexity comparison of the different estimators above is done with respect to exact solutions of the respective SDEs. Of course when Euler approximations are used, the discretization step h must also tend to zero when the required accuracy ε tends to zero. However, it is easy to see that with respect to approximate Euler scheme solutions the same conclusions also can be drawn.

9.3. Numerical experiments

We have implemented the classical and forward–reverse kernel estimator for the one-dimensional example of Section 5. We fix $a = -1$, $b = 1$ and choose fixed initial data $t = 0$, $x = 1$, $T = 1$, $y = 0$, for which $p = 0.518\,831$.

Let us aim to approximate the 'true' value $p = 0.518\,831$ using both the forward–reverse estimator (FRE) and the classical forward estimator (FE). Throughout this experiment we choose $t^* = 0.5$, and $M = N$ for the FRE and the FE is simply obtained by taking $t^* = 1$. For the bandwidth we take $\delta_N^{\text{FE}} = N^{-1/5}$ and $\delta_N^{\text{FRE}} = N^{-1}$, yielding variances $\sigma_{\text{FE}}^2 \approx C_1 N^{-4/5}$ and $\sigma_{\text{FRE}}^2 \approx C_2 N^{-1}$, respectively. It is clear that σ_{FE} may be estimated directly from the density estimation since the classical estimator is proportional to a double sum of N independent random variables. As the FRE is proportional to a double sum of generally *dependent* random variables it is, of course, strictly not correct to estimate its deviation in the same way by just treating these random variables as independent. However, the result of such an (in fact) incorrect estimation, denoted below by σ^*, turns out to be roughly proportional to the correct deviation σ_{FRE}. To show this we estimate σ_{FRE} for $N = 10^2, 10^3, 10^4$, respectively, by running 50 FRE simulations for each value of N and then compute the ratios $\kappa := \sigma_{\text{FRE}}/\sigma^*$ (see Table 1). The SDEs are simulated by the Euler scheme with time step $\Delta t = 0.01$.

So, in general applications we recommend this procedure for determination of the ratio κ which may be carried out with relatively low sample sizes and allows for simple estimation of the variance σ_{FRE}^2. If, for instance, we define the Monte Carlo simulation error to be two standard deviations, the Monte Carlo error of the FRE may be approximated by $2\kappa\sigma^*$.

In this paper we have not addressed the time discretization error due to the numerical scheme used for the simulation of the SDEs. In fact, this is conceptually the same as

Table 1. 'True' σ_{FRE}, σ^* and $k = \sigma_{\text{FRE}}/\sigma^*$ for different N

N	σ_{FRE}	σ^*	κ
10^2	0.068	0.050	1.4
10^3	0.021	0.015	1.4
10^4	0.007	0.005	1.4

Table 2. Estimation of target density p by FRE and FE; 'True' $p = 0.518\,831$

N	FRE	$2\sigma_{\mathrm{FRE}}$	$\sigma^2_{\mathrm{FRE}}N$	Computation time (s)	FE	$2\sigma_{\mathrm{FE}}$	$\sigma^2_{\mathrm{FE}}N^{4/5}$	Computation time (s)
10^4	0.522	0.031	2.40	2	0.524	0.036	0.51	2
10^5	0.519	0.010	2.50	20	0.515	0.016	0.64	18
10^6	0.5194	0.0031	2.45	203	0.5164	0.0064	0.65	183
10^7	0.5193	0.0010	2.50	2085	0.5171	0.0026	0.68	1854

assuming that we have at our disposal a weak numerical scheme of sufficiently high order. We note that if a relatively high accuracy is required in practice, the Euler scheme turns out to be inefficient, as it involves a high number of time steps, yielding, in combination with a high number of paths, a huge complexity. Fortunately, in most cases it will be sufficient to use a weak second-order scheme, such as the method of Talay and Tubaro (1990). The application of this method comes down to Richardson extrapolation of the results obtained by the Euler method for time steps $2\Delta t$ and Δt, respectively. However, we have to take into account that the deviation of this extrapolation, and therefore the Monte Carlo error, is $\sqrt{5}$ times as high. In the experiments below we compare the FRE with the classical one for different sample sizes. For both estimators FRE and FE we use the weak order $\mathcal{O}((\Delta t)^2)$ method of Talay and Tubaro with time discretization steps $\Delta t = 0.02$ and $\Delta t = 0.01$. From Table 2 it is obvious that for larger N the FRE gives a higher Monte Carlo error than the pure FE, while the computational effort involved in the FRE is only a little bit larger. For example, the FRE gives for $N = 10^6$ almost the same Monte Carlo error as the FE for $N = 10^7$. Moreover, due to the choice $\delta_N = N^{-1}$ in the FRE, the bias of the FRE is $\mathcal{O}(N^{-2})$ and so negligible with respect to its deviation being $\mathcal{O}(N^{-1/2})$. Unlike the FRE, with the usual choice $\delta_N = N^{-1/5}$, the bias of the FE is of the same order as its deviation and so its overall error is even larger than its Monte Carlo error displayed in Table 2.

References

Bally, V. and Talay, D. (1996a) The law of the Euler scheme for stochastic differential equations I: Convergence rate of the distribution function. *Probab. Theory Related Fields*, **104**, 43–60.

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

Bretagnolle, J. and Carol-Huber, C. (1979) Estimation des densités: risque minimax. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **47**, 119–137.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: The L_1 View.* New York: Wiley.

Dynkin, E.B. (1965) *Markov Processes.* Berlin: Springer-Verlag.

Greengard, L. and Strain, J. (1991) The fast Gauss transform. *SIAM J. Sci. Statist. Comp.*, **12**(1), 79–94.

Hu, Y. and Watanabe, S. (1996) Donsker delta functions and approximations of heat kernels by the time discretization method. *J. Math. Kyoto Univ.*, **36**, 494–518.

Kohatsu-Higa, A. (1997) High order Ito–Taylor approximations to heat kernels. *J. Math. Kyoto Univ.*, **37**, 129–150.

Kurbanmuradov, O., Rannik, Ü., Sabelfeld, K. and Vesala, T. (1999) Direct and adjoint Monte Carlo for the footprint problem. *Monte Carlo Methods Appl.*, **5**, 85–111.

Kurbanmuradov, O., Rannik, Ü., Sabelfeld, K. and Vesala, T. (2001) Evaluation of mean concentration and fluxes in turbulent flows by Lagrangian stochastic models. *Math. Comput. Simulation*, **54**, 459–476.

Milstein, G.N. (1995) *Numerical Integration of Stochastic Differential Equations*. Dordrecht: Kluwer Academic.

Milstein, G.N. and Schoenmakers, J.G.M. (2002) Monte Carlo construction of hedging strategies against multi-asset European claims. *Stochastics Stochastics Rep.*, **73**, no. 1–2, 125–157.

Newton, N.J. (1994) variance reduction for simulated diffusion. *SIAM J. Appl. Math.*, **54**, 1780–1805.

Pinsker, M.S. (1980) Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission*, **16**, 120–133.

Scott, D.W. (1992) *Multivariate Density Estimation.* New York: Wiley.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Talay, D. and Tubaro, L. (1990) Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, **8**, 483–509.

Thomson, D.J. (1987) Criteria for the selection of stochastic models of particle trajectories in turbulent flows. *J. Fluid. Mech.*, **180**, 529–556.

Wagner, W. (1988) Monte Carlo evaluation of functionals of solutions of stochastic differential equations. Variance reduction and numerical examples. *Stochastic Anal. Appl.*, **6**, 447–468.

Received October 2001 and revised August 2003