Weierstraß–Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

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

G. N. Milstein[1], J. G. M. Schoenmakers[2], V. Spokoiny[2]

submitted: 26th September 2001

[1] Weierstraß Institut
für Angewandte Analysis und Stochastik
Berlin, Germany
and
Ural State University
Ekaterinburg, Russia
E-Mail: milstein@wias-berlin.de

[2] Weierstraß Institut
für Angewandte Analysis und Stochastik
Berlin, Germany
E-Mail: schoenma@wias-berlin.de
E-Mail: spokoiny@wias-berlin.de

Preprint No. 680

Berlin 2001

1991 *Mathematics Subject Classification.* 62G07, 60H10, 65C05.

Key words and phrases. transition density, forward and reverse diffusion, statistical estimation, Monte Carlo simulation.

Edited by
Weierstraß–Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax:	+ 49 30 2044975
E-Mail (X.400):	c=de;a=d400-gw;p=WIAS-BERLIN;s=preprint
E-Mail (Internet):	preprint@wias-berlin.de
World Wide Web:	http://www.wias-berlin.de/

Abstract

The general reverse diffusion equations are derived. They are applied to the problem of transition density estimation of diffusion processes between two fixed states. For this problem it is shown that density estimation based on forward-reverse representations allows for achieving essentially better results in comparison with usual kernel or projection estimation based on forward representations only.

1 Introduction

Consider the SDE in the Ito sense

$$dX = a(s,X)ds + \sigma(s,X)dW(s), \qquad t_0 \leq s \leq T, \tag{1.1}$$

where $X = (X^1, ..., X^d)^\top$, $a = (a^1, ..., a^d)^\top$ are d-dimensional vectors, $W = (W^1, ..., W^m)^\top$ is an m-dimensional standard Wiener process, $\sigma = \{\sigma^{ij}\}$ is a $d \times m$-matrix, $m \geq d$. We assume that the $d \times d$-matrix $b := \sigma\sigma^\top$, $b = \{b^{ij}\}$, is of full rank for every (s,x), $s \in [t_0, T]$, $x \in R^d$. The functions $a^i(s,x)$ and $\sigma^{ij}(s,x)$ are assumed to be sufficiently good in analytical sense (for example, their first derivatives are continuous and bounded). This particularly implies existence and uniqueness of the solution $X_{t,x}(s) \in R^d$, $X_{t,x}(t) = x$, $t_0 \leq t \leq s \leq T$, of (1.1), smoothness of the transition density $p(t,x,s,y)$ of the Markov process X, and existence of all the moments of $p(\cdot,\cdot,\cdot,y)$.

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 realisations of $X_{t,x}(T)$ directly.

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

$$I(f) = \int p(t,x,T,y)f(y)dy. \tag{1.2}$$

These methods exploit the probabilistic representation $I(f) = \boldsymbol{E}\, f(X_{t,x}(T))$. Let $\bar{X}_{t,x}$ be an approximation of the process $X_{t,x}$ and let $\bar{X}_{t,x}^{(n)}(T)$ for $n = 1, \ldots, N$ be independent realizations of $\bar{X}_{t,x}(T)$. Then $I(f)$ may be estimated by

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

with a statistical error of order $N^{-1/2}$, provided the accuracy of approximating $X_{t,x}$ by $\bar{X}_{t,x}$ is sufficiently good.

The problem of estimating the transition density of a diffusion process is more involved, see [1], [6], [7]. For an approximation $\bar{X}_{t,x}$, 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 $\bar{X}_{t,x}(T,h)$ is the approximation of $X_{t,x}(T)$ obtained via numerical integration by the Euler scheme with time step h, then the density $\bar{p}(t,x,T,y)$ converges to $p(t,x,T,y)$ uniformly in y when the step size h tends to zero, see Bally and Talay [2]. Further, in [6] and [7] it is shown that the quantity

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

with $\phi_h(x) = (2\pi h^2)^{-d/2} \exp\{-|x|^2/(2h^2)\}$ converges to $p(t,x,T,y)$ as $h \to 0$. In [6] strong schemes of numerical integration were used, while [7] applied weak schemes. Combining these result with the classical Monte Carlo methods leads to the following estimator of the transition density

$$\widetilde{p}(t,x,T,y) = \frac{1}{N} \sum_{n=1}^{N} \phi_h\left(\bar{X}_{t,x}^{(n)}(T,h) - y\right), \tag{1.3}$$

where $\bar{X}_{t,x}^{(n)}(T,h)$, $n = 1, \ldots, N$, are independent realizations of $\bar{X}_{t,x}(T,h)$. More generally, since the random variables $X_n = \bar{X}_{t,x}^{(n)}(T,h)$ of independent realizations of $\bar{X}_{t,x}(T,h)$ for $n = 1, \ldots, N$ are i.i.d. with the distribution that approximates the distribution of $X_{t,x}(T)$, one may estimate the transition density $p(t,x,T,y)$ from this sample 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

$$\widehat{p}(t,x,T,y) = \frac{1}{N\delta^d} \sum_{n=1}^{N} K\left(\frac{X_n - y}{\delta}\right), \tag{1.4}$$

see e.g. [4]. Clearly, proposal (1.3) is a special case of this estimator with kernel K being the standard normal density and bandwidth δ equal to the step of numerical integration h.

The estimation loss $\widehat{p}(t,x,T,y) - p(t,x,T,y)$ can be split up into an error due to a numerical approximation of the process X by $\bar{X}$ and an error 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 most 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 e.g. [4]). Even an optimal choice of the bandwidth δ leads to quite poor estimation quality, in particular for large dimension d. More specifically, if the underlying density is known to be two times continuously differentiable then the optimal bandwidth δ is of order $N^{-1/(4+d)}$ leading to the accuracy of order $N^{-2/(4+d)}$, see [4]. For $d > 2$, this would require a huge sample size N for providing a reasonable accuracy of estimation. In the statistical literature this problem is referred to as "curse of dimensionality".

Note that the "curse of dimensionality" problem doesn't encounter by the estimation of functionals $I(f)$ in (1.2). Similarly, via probabilistic representations based on reverse diffusion, Monte Carlo estimation of functionals of the form

$$I^*(g) = \int g(x)p(t,x,T,y)dx \tag{1.5}$$

goes with root-N accuracy also, see Section 3. In this paper we aim to propose a method for estimating the transition density $p(t,x,T,y)$ of a diffusion process which allows for root-N consistent estimation for particular values of t, x, T, and y. In this method both the forward and reverse diffusion process are involved.

In Section 2, we discuss some probabilistic representations for the functionals $I(f)$ in (1.2), which thus lead to different Monte-Carlo methods for the evaluating of $I(f)$. Also we show how the error of the Monte Carlo estimation can be reduced by the choice of a suitable probabilistic representation. In Section 3, we introduce the reverse diffusion system in connection with probabilistic representations for functionals of the form (1.5). In Section 4, we explain how the combination of forward and reverse diffusion can be used for efficient Monte Carlo estimation of the transition density. We introduce two different estimators which we refer to as *kernel* and *projection* estimators. General properties of these estimators are studied in Sections 6 and 7. In Section 5 we demonstrate the advantages of combining the forward and reverse diffusion for transition density estimation at a simple one dimensional example. We show by an explicit analysis of an Ornstein-Uhlenbeck type process that root-N accuracy can be achieved. In Section 8 we compare the computational complexity of the forward-reverse kernel estimator with the usual forward kernel estimator 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 only forward diffusion.

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 (1.2). 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

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

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 $\bar{X}$ be an approximation of the process X in (1.1), obtained by some numerical integration scheme. With $\bar{X}_{t,x}^{(n)}(T)$ being independent realizations of $\bar{X}_{t,x}(T)$, the value $u(t,x)$ can be estimated by

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

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

$$I(f,g) = \iint g(x)p(t,x,T,y)f(y)\,dx\,dy.$$

The estimation error $|\widehat{u}-u|$ of the estimator $\widehat{u}$ 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 $\{N^{-1}\,\mathbf{Var}\,f(\bar{X}_{t,x}(T))\}^{1/2} \simeq \{N^{-1}\,\mathbf{Var}\,f(X_{t,x}(T))\}^{1/2}$ and can, in general, be reduced by using variance reduction methods. Variance reduction methods can be derived from the following generalized probabilistic representation for $u(t,x)$:

$$u(t,x)=\boldsymbol{E}\,[f(X_{t,x}(T))\mathcal{X}_{t,x}(T)+\mathbb{X}_{t,x}(T)], \tag{2.3}$$

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

$$\begin{aligned} dX &= (a(s,X)-\sigma(s,X)h(s,X))ds+\sigma(s,X)dW(s), & X(t)&=x,\\ d\mathcal{X} &= h^{\top}(s,X)\mathcal{X}dW(s), & \mathcal{X}(t)&=1,\\ d\mathbb{X} &= F^{\top}(s,X)\mathcal{X}dW(s), & \mathbb{X}(t)&=0. \end{aligned} \tag{2.4}$$

In (2.4), $\mathcal{X}$ and $\mathbb{X}$ are scalars, and $h(t,x)=(h^1(t,x),...,h^m(t,x))^{\top}\in \mathbb{R}^m,\ F(t,x)=(F^1(t,x),...,F^m(t,x))^{\top}\in \mathbb{R}^m$ 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 (2.3)–(2.4) with $h=0,\ F=0$, see, e.g., [5]. The representation for $h\neq 0,\ F=0$ follows from Girsanov's theorem and then we get (2.3) since $\boldsymbol{E}\,\mathbb{X}=0$.

Consider the random variable $\eta := f(X_{t,x}(T))\mathcal{X}_{t,x}(T)+\mathbb{X}_{t,x}(T)$. While the mathematical expectation $\boldsymbol{E}\,\eta$ does not depend on h and F, the variance $\mathbf{Var}\,\eta = \boldsymbol{E}\,\eta^2-(\boldsymbol{E}\,\eta)^2$ does. The Monte Carlo error in (2.2) is of order $\sqrt{N^{-1}\mathbf{Var}\,\eta}$ and so by reduction of the variance $\mathbf{Var}\,\eta$ the Monte Carlo error may be reduced. Two variance reduction methods are well known: the method of importance sampling where $F=0$, see [10], [12], [15], and the method of control variates where $h=0$, see [12]. For both methods it is shown that for sufficiently smooth function f the variance can be reduced to zero. A more general statement is given in Theorem 2.1 below, see also [11]. Introduce the process

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

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 $x \in I\!R^m$ there is a solution of the system (2.4) on the interval $[t,T]$. Then the variance* $\mathbf{Var}\,\eta(T)$ *is equal to*

$$\mathbf{Var}\,\eta(T) = \boldsymbol{E} \int_t^T \mathcal{X}_{t,x}^2(s) \sum_{j=1}^m \left(\sum_{i=1}^d \sigma^{ij} \frac{\partial u}{\partial x^i} + uh^j + F^j \right)^2 ds \tag{2.5}$$

provided that the mathematical expectation in (2.5) exists.

In particular, if h and F satisfy

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

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

Proof. The Ito formula implies

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

and then by $Lu = 0$ we have

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

Hence, (2.5) follows and the last assertion is obvious. □

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

3 Representations relying on reverse diffusion

In the previous section a broad class of probabilistic representations for the integral functionals $I(f) = \int f(y)p(t,x,T,y)dy$, and more generally, for the functionals $I(f,g) = \iint g(x)p(t,y,T,y)f(y)dx\,dy$ is described. Another approach is based on the so called *reverse diffusion* and has been introduced by Thomson [14] (see also [8], [9]). We here 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

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

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

For a given function g and fixed $t>0$ we define

$$v(s,y) := \int g(x')p(t,x',s,y)dx', \qquad s>t,$$

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 $\widetilde{s} = T+t-s$ and define

$$\begin{aligned} \widetilde{v}(\widetilde{s},y) &= v(T+t-\widetilde{s},y), \\ \widetilde{a}^i(\widetilde{s},y) &= a^i(T+t-\widetilde{s},y), \\ \widetilde{b}^{ij}(\widetilde{s},y) &= b^{ij}(T+t-\widetilde{s},y). \end{aligned}$$

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

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

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

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

where

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

So we obtain a Cauchy problem in reverse time and may state the following result.

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

$$I^*(g) = v(T, y) = \widetilde{v}(t, y) = \boldsymbol{E}\left[g(Y_{t,y}(T))\mathcal{Y}_{t,y}(T)\right], \tag{3.4}$$

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

$$\begin{aligned} dY &= \alpha(s, Y)ds + \widetilde{\sigma}(s, Y)d\widetilde{W}(s), \quad & Y(t) = y, \\ d\mathcal{Y} &= c(s, Y)\mathcal{Y}ds, & \mathcal{Y}(t) = 1, \end{aligned} \tag{3.5}$$

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

It is natural to call (3.5) *the reverse system* of (1.1). The probabilistic representation (3.4)–(3.5) for the integral (3.1) leads naturally to the Monte Carlo estimator $\widehat{v}$ for $v(T, y)$,

$$\widehat{v} = \frac{1}{M}\sum_{m=1}^{M} g\left(\bar{Y}_{t,y}^{(m)}(T)\right)\bar{\mathcal{Y}}_{t,y}^{(m)}(T), \tag{3.6}$$

where $(\bar{Y}_{t,y}^{(m)}, \bar{\mathcal{Y}}_{t,y}^{(m)})$, $m = 1, \ldots, M$, are independent realizations of the process $(\bar{Y}_{t,y}, \bar{\mathcal{Y}}_{t,y})$ that approximates the process $(Y_{t,y}, \mathcal{Y}_{t,y})$ from (3.5).

Similar to (2.3)–(2.4), the representation (3.4)–(3.5) may be extended to

$$v(T, y) = \boldsymbol{E}\left[g(Y_{t,y}(T))\mathcal{Y}_{t,y}(T) + \mathbb{Y}_{t,y}(T)\right], \tag{3.7}$$

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

$$\begin{aligned} dY &= (\alpha(s, Y) - \widetilde{\sigma}(s, Y)\widetilde{h}(s, Y))ds + \widetilde{\sigma}(s, Y)d\widetilde{W}(s), \quad & Y(t) = y, \\ d\mathcal{Y} &= c(s, Y)\mathcal{Y}ds + \widetilde{h}^\top(s, Y)\mathcal{Y}d\widetilde{W}(s), & \mathcal{Y}(t) = 1, \\ d\mathbb{Y} &= \widetilde{F}^\top(s, Y)\mathcal{Y}d\widetilde{W}(s), & \mathbb{Y}(t) = 0. \end{aligned} \tag{3.8}$$

In (3.8), $\mathcal{Y}$ and $\mathbb{Y}$ are scalars, $\widetilde{h}(t, x) \in I\!R^m$, and $\widetilde{F}(t, x) \in I\!R^m$ are arbitrary vector functions which satisfy some regularity conditions.

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

Remark 3.2. By constructing the reverse system of reverse system (3.5), we get the original system (1.1) accompanied by a scalar equation with coefficient $-c$. By then taking the reverse of this system we get (3.5) again.

Remark 3.3. If the original stochastic system (1.1) is linear, then the system (3.5) 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 a probabilistic representation for the target probability density $p(t,x,T,y)$, which utilizes both the forward and the reverse diffusion system. Next, we give two different Monte Carlo estimators for $p(t,x,T,y)$ based on this representation: a kernel estimator and a 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_1 be an internal point of the interval $[t,T]$. By the Kolmogorov-Chapman equation for the transition density we have

$$p(t,x,T,y) = \int p(t,x,t_1,x')p(t_1,x',T,y)dx'. \tag{4.1}$$

By applying Theorem 3.1 with $g(x') = p(t,x,t_1,x')$, it follows that this equation has a probabilistic representation,

$$p(t,x,T,y) = \boldsymbol{E}\, p(t,x,t_1,Y_{t_1,y}(T))\,\mathcal{Y}_{t_1,y}(T). \tag{4.2}$$

Since in general the density function $x' \to p(t,x,t_1,x')$ is unknown also, we cannot apply the Monte Carlo estimator $\widehat{u}$ in (2.2) 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 $[t,t_1]$ by standard methods of non-parametric statistics and then to replace in the r.h.s. of (4.2) the unknown density function by its estimator, say $x' \to \widehat{p}(t,x,t_1,x')$. This idea suggests the following procedure. Generate by numerical integration of the forward system (1.1) and the reverse system (3.5) (or (3.8)) independent samples $\bar{X}^{(n)}_{t,x}(t_1)$, $n = 1,\ldots,N$ and $(\bar{Y}^{(m)}_{t_1,y}(T), \bar{\mathcal{Y}}^{(m)}_{t_1,y}(T))$, $m = 1,\ldots,M$, respectively (in general different step sizes may be used for $\bar{X}$ and $\bar{Y}$). Let $\widehat{p}(t,x,t_1,x')$ be, for instance, the kernel estimator of $p(t,x,t_1,x')$ from (1.4), that is,

$$\widehat{p}(t,x,t_1,x') = \frac{1}{N\delta^d}\sum_{n=1}^{N} K\left(\frac{\bar{X}_{t,x}^{(n)}(t_1)-x'}{\delta}\right).$$

Thus, replacing p by this kernel estimator in the r.h.s. of (4.2) yields a forward representation of the form (2.1) which in turn may be estimated by

$$\widehat{p}(t,x,T,y) = \frac{1}{M}\left[\frac{1}{N\delta_N^d}\sum_{m=1}^{M}\sum_{n=1}^{N} K\left(\frac{\bar{X}_{t,x}^{(n)}(t_1)-\bar{Y}_{t_1,y}^{(m)}(T)}{\delta_N}\right)\bar{\mathcal{Y}}_{t_1,y}^{(m)}(T)\right]. \tag{4.3}$$

We will show that this heuristic idea really works and leads to estimators which have superior properties in comparison with usual density estimators based on pure forward or pure reverse representations. Of course, the kernel estimation of $p(t,x,t_1,x')$ in the first step will be crude as usual for a particular x'. But, due to a good overall property of kernel estimators, namely, the fact that any kernel estimator is a density, the impact of these point-wise errors will be reduced in the second step, the estimation of (4.2). In fact, by the Kolmogorov-Chapman 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 achievable accuracy 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*

$$\begin{aligned} J(f) &:= \iint p(t,x,t_1,x')p(t_1,y',T,y)f(x',y')dx'dy' \\ &= \boldsymbol{E}\left[f(X_{t,x}(t_1),Y_{t_1,y}(T))\,\mathcal{Y}_{t_1,y}(T)\right], \end{aligned} \tag{4.4}$$

where $X_{t,x}(s)$ obeys the forward equation (1.1) and $(Y_{t_1,y}(s),\mathcal{Y}_{t_1,y}(s))$, $s\geq t_1$, is the solution of the reverse system (3.5).

Proof. Conditioning on $X_{t,x}(t_1)$ and applying Theorem 3.1 with $g(\cdot)=f(x',\cdot)$ for every x' yields

$$\begin{aligned} \boldsymbol{E}\left(f(X_{t,x}(t_1),Y_{t_1,y}(T))\,\mathcal{Y}_{t_1,y}(T)\right) &= \boldsymbol{E}\,\boldsymbol{E}\left(f(X_{t,x}(t_1),Y_{t_1,y}(T))\,\mathcal{Y}_{t_1,y}(T)\mid X_{t,x}(t_1)\right) \\ &= \int p(t,x,t_1,x')\left(\int f(x',y')p(t_1,y',T,y)dy'\right)dx'. \end{aligned}$$

□

Let $\bar{X}^{(n)}_{t,x}(t_1)$, $n = 1, \ldots, N$, be a sample of independent realizations of an approximation $\bar{X}$ of X, obtained by numerical integration of (1.1) on the interval $[t, t_1]$. Similarly, let $(\bar{Y}^{(m)}_{t_1,y}(T)\bar{\mathcal{Y}}^{(m)}_{t_1,y}(T))$, $m = 1, \ldots, M$ be independent realizations of a numerical solution of (3.5) on the interval $[t_1, T]$. Then the representation (4.4) leads to the following Monte Carlo estimator for $J(f)$,

$$\widehat{J} = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} f\left(\bar{X}^{(n)}_{t,x}(t_1), \bar{Y}^{(m)}_{t_1,y}(T)\right) \bar{\mathcal{Y}}^{(m)}_{t_1,y}(T). \tag{4.5}$$

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

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

where $\delta^{-d}K(u/\delta)$ converge to $\delta_0(u)$ (in distribution sense) as $\delta \downarrow 0$. Then the corresponding expression for $\widehat{J}$ coincides with the kernel estimator $\widehat{p}$ 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 $\{\varphi_\ell\,,\, \ell \geq 1\}$ is a total orthonormal system in the function space $L_2(\mathbb{R}^d)$ and L is a natural number. It is known that $f_{\varphi,L} \to \delta_{\text{diag}}$ (in distribution sense) as $L \to \infty$. This leads to the projection estimator,

$$\widehat{p}^{pr} = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{\ell=1}^{L} \varphi_\ell\left(\bar{X}^{(n)}_{t,x}(t_1)\right) \varphi_\ell\left(\bar{Y}^{(m)}_{t_1,y}(T)\right) \bar{\mathcal{Y}}^{(m)}_{t_1,y}(T) = \sum_{\ell=1}^{L} \widehat{\alpha}_\ell\, \widehat{\gamma}_\ell, \tag{4.6}$$

with

$$\widehat{\alpha}_\ell = \frac{1}{N} \sum_{n=1}^{N} \varphi_\ell\left(\bar{X}^{(n)}_{t,x}(t_1)\right), \qquad \widehat{\gamma}_\ell = \frac{1}{M} \sum_{m=1}^{M} \varphi_\ell\left(\bar{Y}^{(m)}_{t_1,y}(T)\right) \bar{\mathcal{Y}}^{(m)}_{t_1,y}(T).$$

The general properties of the kernel estimator are studied in Section 6 and the projection estimator is studied in Section 7. As mentioned previously, by selecting properly a weak scheme and step size h, approximate solutions of systems of SDEs can be simulated sufficiently close to exact solutions. Therefore, in what follows we do not distinguish between the process $X_{t,x}(s)$, respectively $(Y_{t_1,y}(s), \mathcal{Y}_{t_1,y}(s))$, and their approximation $\bar{X}_{t,x}(s)$, respectively $(\bar{Y}_{t_1,y}(s), \bar{\mathcal{Y}}_{t_1,y}(s))$. Moreover, by skipping these not really essential technicalities we may keep our exposition more transparent.

Remark 4.1. In general it is possible to apply variance reduction methods to the estimator $\widehat{J}$ in (4.5), based on the extended representations (2.3)–(2.4) and (3.7)–(3.8).

5 The 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

$$dX = aX dt + bdW(t), \qquad X(0) = x, \tag{5.1}$$

which is known for $a < 0$ as the Ornstein-Uhlenbeck process. By (3.5), the reverse system belonging to (5.1) is given by

$$\begin{aligned} dY &= -aY ds + bd\widetilde{W}(s), \ Y(t) = y, \ s > t, && (5.2)\\ d\mathcal{Y} &= -a\mathcal{Y} ds, \ \mathcal{Y}(t) = 1. && (5.3) \end{aligned}$$

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

$$X(t) = e^{at}(x + b\int_0^t e^{-au} dW(u))$$

and

$$\begin{aligned} Y(s) &= e^{-a(s-t)}(y + b\int_t^s e^{a(u-t)} d\widetilde{W}(u)),\\ \mathcal{Y}(s) &= e^{-a(s-t)}, \end{aligned}$$

respectively. It follows that

$$\boldsymbol{E}\, X(t) = e^{at}x, \quad \mathbf{Var}\, X(t) = b^2 e^{2at}\int_0^t e^{-2au} du = b^2 \frac{e^{2at}-1}{2a} := \sigma^2(t)$$

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

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

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

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

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 $0 \leq t_1 \leq T$. For simplicity, we don't deal with variance reduction, i.e, we take $h \equiv 0$ and $F \equiv 0$. It follows that

$$p_X(0,x,T,y) \simeq \xi_{N,M} := \frac{e^{-a(T-t_1)}}{MN\delta} \sum_{m=1}^{M} \sum_{n=1}^{N} K_{nm}, \tag{5.5}$$

where

$$\begin{aligned} K_{nm} &:= K((e^{at_1}(x + b\int_0^{t_1} e^{-au} dW^{(n)}(u)) - e^{-a(T-t_1)}(y + b\int_{t_1}^{T} e^{a(u-t_1)} d\widetilde{W}^{(m)}(u)))\delta^{-1}) \\ &= K((e^{at_1}x - e^{-a(T-t_1)}y + \sigma(t_1)U^{(n)} - e^{-a(T-t_1)}\sigma(T-t_1)V^{(m)})\delta^{-1}) \end{aligned} \tag{5.6}$$

with $U^{(n)}$ and $V^{(m)}$ being i.i.d. standard normally distributed random variables. Note that in general δ in (5.5) and (5.6) may be chosen in dependence of both N and M, so $\delta = \delta_{N,M}$ in fact. It is clear that (5.5) collapses to a classical (pure) forward estimator or (pure) reverse estimator if $t_1 = 0$, or $t_1 = T$, respectively.

By choosing the Gaussian kernel

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

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

$$\boldsymbol{E}\,\xi_{N,M} = \frac{1}{\sqrt{2\pi\,(\delta^2 e^{2a(T-t_1)} + \sigma^2(T))}} \exp[-\frac{(e^{aT}x - y)^2}{2(\delta^2 e^{2a(T-t_1)} + \sigma^2(T))}] \tag{5.8}$$

and for the variance it follows that

$$\begin{aligned} \mathbf{Var}\,(\xi_{N,M}) &= \tfrac{-N-M+1}{2\pi MN(B+\sigma^2(T))} \exp[-\tfrac{A}{B+\sigma^2(T)}] \\ &+ \tfrac{M-1}{2\pi MN\sqrt{B+\sigma^2(T-t_1)}\sqrt{B+2\sigma^2(T)-\sigma^2(T-t_1)}} \exp[-\tfrac{A}{B+2\sigma^2(T)-\sigma^2(T-t_1)}] \\ &+ \tfrac{N-1}{2\pi MN\sqrt{B+\sigma^2(T)-\sigma^2(T-t_1)}\sqrt{B+\sigma^2(T)+\sigma^2(T-t_1)}} \exp[-\tfrac{A}{B+\sigma^2(T)+\sigma^2(T-t_1)}] \\ &+ \tfrac{e^{-a(T-t_1)}}{2\pi MN\delta\sqrt{B+2\sigma^2(T)}} \exp[-\tfrac{A}{B+2\sigma^2(T)}]. \end{aligned} \tag{5.9}$$

with the abbreviations $A := (e^{aT}x - y)^2$, $B := \delta^2 e^{2a(T-t_1)}$. Since in Sections 6 the forward reverse kernel estimator will be analysed quite general, we here sketch the derivation of (5.8) and (5.9) just briefly. It is convenient to use the following standard lemma which we state without proof.

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

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

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

$$\boldsymbol{E}\,\xi_{N,M}^2 = \frac{e^{-2a(T-t_1)}}{M^2N^2\delta^2}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{m'=1}^{M}\sum_{n'=1}^{N}\boldsymbol{E}\,K_{nm}K_{n'm'} \tag{5.10}$$

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 $\mathbf{Var}\,(\xi_{N,M}) = \boldsymbol{E}\,\xi_{N,M}^2 - (\boldsymbol{E}\,\xi_{N,M})^2$.

We now compare the bias and variance of (5.5) for $0 < t_1 < T$ with the classical cases $t_1 = 0$ and $t_1 = T$. The bias in (5.8) converges to zero for $\delta \downarrow 0$, since we have

$$\boldsymbol{E}\,\xi_{N,M} = \frac{\exp[-\frac{(e^{aT}x-y)^2}{2\sigma^2(T)}]}{\sqrt{2\pi\sigma^2(T)}}(1+\mathcal{O}(\delta^2)) = p_X(0,x,T,y)(1+\mathcal{O}(\delta^2)).$$

So, the bias is of order $\mathcal{O}(\delta^2)$ and thus the same as in the classical situation for a kernel given by (5.7). For $t_1 = T$ we obtain the classical pure forward estimator and by substituting $t_1 = T$ in (5.9) we get the variance of the classical forward estimator,

$$\mathbf{Var}\,(\xi_{N,M}^{t_1=T}) = \frac{1}{2\pi N}\frac{\exp[-\frac{\left(e^{aT}x-y\right)^2}{\delta^2+2\sigma^2(T)}]}{\delta\sqrt{\delta^2+2\sigma^2(T)}} - \frac{1}{2\pi N}\frac{\exp[-\frac{\left(e^{aT}x-y\right)^2}{\delta^2+\sigma^2(T)}]}{\delta^2+\sigma^2(T)}, \tag{5.11}$$

where M has dropped out since there is no reverse simulation in fact. Similarly, for $t_1 = 0$ we obtain the classical reverse estimator with variance

$$\mathbf{Var}\,(\xi_{N,M}^{t_1=0}) = \frac{e^{-aT}}{2\pi M}\frac{\exp[-\frac{\left(e^{aT}x-y\right)^2}{\delta^2e^{2aT}+2\sigma^2(T)}]}{\delta\sqrt{\delta^2e^{2aT}+2\sigma^2(T)}} - \frac{1}{2\pi M}\frac{\exp[-\frac{\left(e^{aT}x-y\right)^2}{\delta^2e^{2aT}+\sigma^2(T)}]}{\delta^2e^{2aT}+\sigma^2(T)}, \tag{5.12}$$

where now N has dropped out since we have only backward simulation. Now, comparison of (5.9) with (5.11) or (5.12) leads to the following interesting conclusion.

Conclusion 5.1. We consider the case $M = N$ and denote the estimator for $p_X(0, x, T, y)$ by ξ_N. The width δ will thus be chosen in relation to N, hence $\delta = \delta_N$. We observe that

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

where $\varepsilon_N := \sqrt{\mathbf{E}(\xi_N - p_X(0,x,T,y))^2}$ is usually referred to as the accuracy of the estimation. From (5.13), (5.11) and (5.12) it is clear that for both pure forward and pure reverse simulation ($t_1 = T$ or $t_1 = 0$, respectively) we have $\varepsilon_N \downarrow 0$ when $N \to \infty$, if and only if $\delta_N \to 0$ *and* $N\delta_N \to \infty$. So, by (5.11) and (5.12) again we have for the classical (forward or reverse) estimator

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

for some positive constants c_1, c_2. It thus follows that the best achievable accuracy rate for the classical estimators 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 $0 < t_1 < T$. From (5.9) and (5.13) it follows by similar arguments that

$$\varepsilon_N^2 = (\frac{d_1}{N} + \frac{d_2}{N^2\delta_N} + d_3\delta_N^4)(1 + o(1)), \quad N\delta_N^2 \to \infty \ \ and \ \ \delta_N \downarrow 0, \tag{5.14}$$

for some positive constants d_1, d_2 and d_3. So, from (5.14) 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 $p \in [\frac{1}{4}, 1]$!

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. Let $r(u)$ be the density of the random variable $X_{t,x}(t_1)$, that is, $r(u) = p(t, x, t_1, u)$. Similarly, let $q(u)$ be the density of $Y_{t_1,y}(T)$ and further denote by $\mu(u)$ the conditional mean of $\mathcal{Y}_{t_1,y}(T)$ given $Y_{t_1,y}(T) = u$. By the following lemma we may reformulate the representation for p in (4.2) and $J(f)$ in (4.4).

Lemma 6.1.

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

$$J(f) = \int f(u,v)r(u)q(v)\mu(v)\,du\,dv. \tag{6.2}$$

Proof. (6.1) follows from (4.2) by

$$\begin{aligned} p &= \boldsymbol{E}\, r\left(Y_{t_1,y}(T)\right)\mathcal{Y}_{t_1,y}(T) = \boldsymbol{E}\left[r\left(Y_{t_1,y}(T)\right)\boldsymbol{E}\left(\mathcal{Y}_{t_1,y}(T) \mid Y_{t_1,y}(T)\right)\right] \\ &= \boldsymbol{E}\, r\left(Y_{t_1,y}(T)\right)\mu(Y_{t_1,y}(T)) = \int r(u)\mu(u)q(u)du \end{aligned} \tag{6.3}$$

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

For a kernel function $K(z)$ in $I\!R^d$ 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,

$$J(f_{K,\delta}) = \int\int \delta^{-d}K(\frac{u-v}{\delta})r(u)q(v)\mu(v)\,du\,dv,$$

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

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

with

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

where $X_n := X_{t,x}^{(n)}(t_1) \in I\!R^d$, $n = 1,\ldots,N$, may be regarded as an i.i.d. sample from the distribution with density r, the sequence $Y_m = Y_{t_1,y}^{(m)}(T) \in I\!R^d$, $m = 1,\ldots,M$, as an i.i.d. sample from the distribution with the density q, and the weights $\mathcal{Y}_m = \mathcal{Y}_{t_1,y}^{(m)}(T)$, $m = 1,\ldots,M$, may be seen as independent samples from a distribution conditional on Y_m, with conditional mean $\mu(y)$ given $Y_m = u$. Below we derive some properties of this estimator.

Lemma 6.2. *We have*

$$\boldsymbol{E}\,\widehat{p} = p_\delta := \int\int r(u+\delta v)q(u)\mu(u)K(v)\,du\,dv = \int r_\delta(u)\lambda(u)du$$

with

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

and

$$r_\delta(u) := \delta^{-d}\int r(v)K\left(\delta^{-1}(v-u)\right)dv = \int r(u+\delta v)K(v)dv.$$

Moreover, if the kernel K fulfills $\int K(u)du = 1$, $K(u)\ge 0$, $K(u) = K(-u)$ for all $u\in I\!R^d$, and $K(u) = 0$ for $|u|>1$, then the bias $|p - \boldsymbol{E}\,\widehat{p}|$ satisfies

$$|p - \boldsymbol{E}\,\widehat{p}| = |p - p_\delta| \le C_K\|r''\|\delta^2 \tag{6.5}$$

with $C_K = \dfrac{1}{2}\int |v|^2K(v)dv \cdot \int \lambda(u)du$ and $\|r''\| = \sup_v\|r''(v)\|$, where $\|r''(v)\|$ is the Euclidean norm of the matrix $r''(v) = \left\{\dfrac{\partial^2 r}{\partial v^i\partial v^j}\right\}$.

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

$$\begin{aligned}\boldsymbol{E}\,Z_{nm} &= \delta^{-d}\int\int r(u)q(v)\mu(v)K\left(\delta^{-1}(u-v)\right)du\,dv\\ &= \int\int r(u+\delta v)q(u)\mu(u)K(v)\,du\,dv = p_\delta\,.\end{aligned}$$

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

$$\begin{aligned}r_\delta(u) - r(u) &= \int r(u+\delta v)K(v)\,dv - r(u) = \int\left[r(u+\delta v) - r(u) - \delta v^\top r'(u)\right]K(v)dv\\ &= \int \frac{1}{2}\delta^2 v^\top r''(u+\theta(v)\delta v)v\,K(v)dv\\ &\le \frac{1}{2}\delta^2\|r''\|\int |v|^2K(v)dv,\end{aligned}$$

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

$$|p_\delta - p| \le \int |r_\delta(u) - r(u)|\lambda(u)du \le C_K\delta^2\|r''\|\int\lambda(u)du.$$

$\square$

Remark 6.1. The order of the bias $|p_\delta - p|$ can be improved by using higher-order kernels for K. We say that K is of order β if it holds $\int u_1^{j_1}\ldots u_d^{j_d}K(u)du = 0$ for all nonnegative integer numbers $j_1,\ldots,j_d$ satisfying $0 < j_1+\ldots+j_d \le \beta$. Similar to the proof of Lemma 6.2 one can show that the application of a kernel K of order β satisfying $\int K(u)du = 1$, $K(u) = 0$ for $|u|\ge 1$, leads to a bias with $|p_\delta - p| \le C\delta^{\beta+1}$, where C is some constant depending on r, q and K.

Concerning the variance $\mathbf{Var}\,\widehat{p} = \boldsymbol{E}\,(\widehat{p} - \boldsymbol{E}\,\widehat{p})^2$ of the estimator (6.4) we obtain the next result.

Lemma 6.3. *It holds*

$$\mathbf{Var}\,\widehat{p} = \frac{1}{NM}\delta^{-d}B_\delta + \frac{M-1}{NM}\int r(u)\lambda_\delta^2(u)du + \frac{N-1}{NM}\int r_\delta^2(u)\mu_2(u)q(u)\,du \\ -\frac{N+M-1}{NM}p_\delta^2\,,$$

where

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

with

$$\begin{aligned} \lambda_\delta(u) &= \delta^{-d}\int \lambda(v)K\left(\delta^{-1}(v-u)\right)dv = \int \lambda(u+\delta v)K(v)dv, \\ r_{\delta,2}(u) &= \delta^{-d}\int r(v)K^2\left(\delta^{-1}(v-u)\right)dv = \int r(u+\delta v)K^2(v)dv, \\ \mu_2(v) &= \boldsymbol{E}\,(\mathcal{Y}_1^2 \mid Y_1 = v). \end{aligned}$$

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

$$\begin{aligned} M^2N^2\mathbf{Var}\,\widehat{p} &= \boldsymbol{E}\left(\sum_{n=1}^{N}\sum_{m=1}^{M}(Z_{nm}-p_\delta)\right)^2 \qquad (6.6) \\ &= \sum_{n=1}^{N}\sum_{m=1}^{M}\boldsymbol{E}\,(Z_{nm}-p_\delta)^2 + \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{m'\neq m}(\boldsymbol{E}\,Z_{nm}Z_{nm'} - p_\delta^2) \\ &\quad + \sum_{n=1}^{N}\sum_{n'\neq n}\sum_{m=1}^{M}(\boldsymbol{E}\,Z_{nm}Z_{n'm} - p_\delta^2). \end{aligned}$$

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

$$\begin{aligned} \boldsymbol{E}\,Z_{nm}Z_{nm'} &= \delta^{-2d}\int\int\int K\left(\delta^{-1}(u-v)\right)K\left(\delta^{-1}(u-v')\right)r(u)\lambda(v)\lambda(v')du\,dv\,dv' \\ &= \delta^{-d}\int\int K\left(\delta^{-1}(u-v)\right)r(u)\lambda_\delta(u)\lambda(v)\,du\,dv \\ &= \int r(u)\lambda_\delta^2(u)\,du \end{aligned}$$

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

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

Further,

$$\begin{aligned}
\boldsymbol{E}\, Z_{nm}^2 &= \delta^{-2d} \boldsymbol{E}\, \mathcal{Y}_m^2 K^2\left(\delta^{-1}\left(X_n - Y_m\right)\right) \\
&= \delta^{-2d} \boldsymbol{E}\, \left(K^2\left(\delta^{-1}\left(X_n - Y_m\right)\right) \boldsymbol{E}\, \left(\mathcal{Y}_m^2 \mid Y_m\right)\right) \\
&= \delta^{-2d} \int\int K^2\left(\delta^{-1}(u-v)\right) r(u) q(v) \mu_2(v)\, du\, dv \\
&= \delta^{-d} \int \mu_2(v) q(v) r_{\delta,2}(v) dv
\end{aligned}$$

and so we get

$$\mathbf{Var}\, \widehat{p} = \frac{\delta^{-d} v_\delta - p_\delta^2}{NM} + \frac{M-1}{NM}\left(\int r(u)\lambda_\delta^2(u) du - p_\delta^2\right) + \frac{N-1}{NM}\left(\int r_\delta^2(u)\mu_2(u)\, q(u) du - p_\delta^2\right)$$

from which the assertion follows. □

Let us define

$$B = \int K^2(u) du \cdot \int r(u)\mu_2(u)\, q(u) du.$$

By the Taylor expansion

$$r(u+\delta v) = r(u) + \delta v^\top r'(u) + \frac{1}{2}\delta^2 v^\top r''(u + \theta(v)\delta v) v,$$

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 get

$$\begin{aligned}
\left|\int r(u)\lambda_\delta^2(u) du - \int r(u)\lambda^2(u) du\right| &= O(\delta^2), \qquad \delta \downarrow 0, \\
\left|\int r_\delta^2(u)\mu_2(u)\, q(u)\, du - \int r^2(u)\mu_2(u)\, q(u)\, du\right| &= O(\delta^2), \qquad \delta \downarrow 0.
\end{aligned}$$

Further, introduce the constant D by

$$D := \int r(u)\lambda^2(u) du + \int r^2(u)\mu_2(u)\, q(u)\, du - 2p^2.$$

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

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

$$\left|\mathbf{Var}\, \widehat{p} - \frac{D}{N} - \frac{\delta^{-d} B}{N^2}\right| \le C\left(\frac{\delta^{-d+2}}{N^2} + \frac{\delta^2}{N} + \frac{1}{N^2}\right). \tag{6.7}$$

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

$$\left|\mathbf{Var}\, \widehat{p} - \frac{D}{N}\right| = \frac{o(1)}{N}, \qquad N \to \infty.$$

Now, by combining Lemmas 6.1 and 6.4 we have the following theorem.

Theorem 6.1. *Let $N = M$ and $\delta = \delta_N$ depend on N. The following statements hold:*

1) If $d < 4$ and δ_N is such that

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

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

$$\boldsymbol{E}\,(\widehat{p} - p)^2 = (p_\delta - p)^2 + \mathbf{Var}\,\widehat{p} = \frac{D}{N} + \frac{o(1)}{N}, \qquad N \to \infty. \tag{6.8}$$

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

2) 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}.

3) 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 (6.7) imply (6.8). 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 really possible to achieve an estimation accuracy of rate $N^{-1/2}$ for $d \leq 4$. Moreover, for $d > 4$ an accuracy rate of root-N may be achieved as well by applying a higher order kernel K.

In section 8 we will see that with the proposed choice of the bandwidth $\delta_N = N^{-1/d}\log^{1/d} N$ for $d \leq 3$ and $\delta_N = N^{-2/(4+d)}$ for $d \geq 4$, the kernel estimator $\widehat{p}$ can be computed at a cost of order $N \log N$ operations.

7 The forward reverse projection estimator

In this section we discuss statistical properties of the *projection* estimator $\widehat{p}^{pr}$ 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 $L_2(I\!R^d)$. For example, in the case $d = 1$ one could take

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

where $H_l(u),\ l \geq 0$, 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 $r(u) = p(t,x,t_1,u)$ (see Section 6) and $h(u) := p(t_1,u,T,y)$ the Fourier expansions

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

$$\alpha_\ell := \int r(u)\varphi_\ell(u)du, \qquad \gamma_\ell := \int h(u)\varphi_\ell(u)du.$$

By (2.1), (3.1), and (3.4) it follows that

$$\alpha_\ell = \boldsymbol{E}\, \varphi_\ell(X_{t,x}(t_1)), \tag{7.1}$$

$$\gamma_\ell = \boldsymbol{E}\, \varphi_\ell(Y_{t_1,y}(T))\mathcal{Y}_{t_1,y}(T), \tag{7.2}$$

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

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

Therefore, it is natural to consider the estimator

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

where L is a natural number and

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

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) coincides 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 analytical sense such that, in particular, the functions $y' \to p(t, x, t_1, y')$ and $y' \to p(t_1, y', T, y)$ are squared integrable. Hence, we assume that the Fourier expansions used in this section are valid in $L_2(I\!R^d)$. The notation introduced in Section 6 is maintained below. We have the following lemma.

Lemma 7.1. *It holds for every* $\ell \geq 1$

$$\begin{aligned} \boldsymbol{E}\,\widehat{\alpha}_\ell &= \alpha_\ell = \int r(u)\varphi_\ell(u)du, \\ \mathbf{Var}\,\widehat{\alpha}_\ell &= N^{-1}\,\mathbf{Var}\,\varphi_\ell(X_1) = N^{-1}\left(\int \varphi_\ell^2(u)r(u)du - \alpha_\ell^2\right) =: N^{-1}\alpha_{\ell,2}. \end{aligned}$$

Similarly,

$$\begin{aligned} \boldsymbol{E}\,\widehat{\gamma}_\ell &= \gamma_\ell = \int \varphi_\ell(u)\mu(u)q(u)du, \\ \mathbf{Var}\,\widehat{\gamma}_\ell &= M^{-1}\,\mathbf{Var}\,\mathcal{Y}_1\varphi_\ell(Y_1) = M^{-1}\left(\int \mu_2(u)\varphi_\ell^2(u)q(u)du - \gamma_\ell^2\right) =: M^{-1}\gamma_{\ell,2}, \end{aligned}$$

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

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 $\widehat{\alpha}_\ell$ and the $\widehat{\gamma}_\ell$'s are independent, it follows by Lemma 7.1 that

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

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

Lemma 7.2. *It holds*

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

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

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

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

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

Then we have

$$N\,\mathbf{Var}\,\widehat{p}^{pr} \le C$$

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

Proof. Let us write

$$\begin{aligned}\sum_{\ell=1}^{L} \widehat{\alpha}_\ell \widehat{\gamma}_\ell - \sum_{\ell=1}^{L} \alpha_\ell \gamma_\ell \;&=\; \sum_{\ell=1}^{L} (\widehat{\alpha}_\ell - \alpha_\ell)(\widehat{\gamma}_\ell - \gamma_\ell) + \sum_{\ell=1}^{L} \alpha_\ell (\widehat{\gamma}_\ell - \gamma_\ell) + \sum_{\ell=1}^{L} (\widehat{\alpha}_\ell - \alpha_\ell)\gamma_\ell \\ &=:\; I_1 + I_2 + I_3.\end{aligned}$$

The Cauchy-Schwarz inequality implies

$$\begin{aligned}\boldsymbol{E}\,(I_2)^2 = \boldsymbol{E}\left(\sum_{\ell=1}^{L} \alpha_\ell(\widehat{\gamma}_\ell - \gamma_\ell)\right)^2 \;&\le\; \boldsymbol{E}\left(\sum_{\ell=1}^{L} |\alpha_\ell| \sum_{\ell=1}^{L} |\alpha_\ell| (\widehat{\gamma}_\ell - \gamma_\ell)^2\right) \\ &\le\; C_{1,\alpha} \sum_{\ell=1}^{L} |\alpha_\ell| \boldsymbol{E}\,(\widehat{\gamma}_\ell - \gamma_\ell)^2 \le C_{1,\alpha}^2 C_{2,\gamma} N^{-1}\end{aligned}$$

and similarly

$$\boldsymbol{E}\,(I_3)^2 = \boldsymbol{E}\left(\sum_{\ell=1}^{L} \gamma_\ell(\widehat{\alpha}_\ell - \alpha_\ell)\right)^2 \le C_{1,\gamma}^2 C_{2,\alpha} N^{-1}.$$

The Cauchy-Schwarz inequality and independence of the $\widehat{\alpha}_\ell$'s and the $\widehat{\gamma}_\ell$'s imply

$$\begin{aligned}\boldsymbol{E}\,(I_1)^2 = \boldsymbol{E}\left(\sum_{\ell=1}^{L} (\widehat{\alpha}_\ell - \alpha_\ell)(\widehat{\gamma}_\ell - \gamma_\ell)\right)^2 \;&\le\; \boldsymbol{E} \sum_{\ell=1}^{L} (\widehat{\alpha}_\ell - \alpha_\ell)^2 \boldsymbol{E} \sum_{\ell=1}^{L} (\widehat{\gamma}_\ell - \gamma_\ell)^2 \\ &\le\; C_{2,\alpha} C_{2,\gamma} (L+1)^2 N^{-2} \le C_{2,\alpha} C_{2,\gamma} N^{-1}.\end{aligned}$$

Hence,

$$\mathbf{Var}\,\widehat{p}^{pr} = \boldsymbol{E}\,(I_1 + I_2 + I_3)^2 \le (\sqrt{E(I_1)^2} + \sqrt{E(I_2)^2} + \sqrt{E(I_3)^2})^2 \le \frac{C}{N}$$

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

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*

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

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

$$\boldsymbol{E}\,(\widehat{p}^{pr} - p)^2 \le CN^{-1}.$$

Proof. Clearly,

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

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

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

Next,

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

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

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

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 usual functional basis such as the Hermite bases, condition (7.8) is fulfilled if the underlying densities $p(t, x, t_1, x')$ and $p(t_1, x', T, y)$ 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 as 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 (e.g. Hermite basis functions) condition (7.8) is fulfilled for any $\beta > d/2$ when the densities $p(t, x, t_1, x')$ and $p(t_1, x', T, y)$ 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 desirable root-N consistency. If, moreover, these densities are analytical one can proof that even $L_N = \mathcal{O}(\log N)$ leads to root-N consistency. Generally it is clear that properly choosing L_N is essential for reducing the numerical complexity of the procedure, see Section 8.

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 transferred into conditions on the coefficients of the original diffusion model (1.1) and the chosen orthonormal basis. Note, however, that in the case of e.g. 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 $\sigma(s,x)\sigma^\top(s,x)$ is of full rank for all s, x.

8 Implementation of the forward-reverse estimators, complexity of the estimation algorithms, numerical examples

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 with 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 the cost of MN kernel evaluations which would be tremendous, for example, when $M = N = 10^5$! But, fortunately, by using kernels with an in some sense small support we can get around this difficulty as outlined below.

Implementation of the kernel estimator and its numerical complexity

We here assume that the kernel $K(x)$ used in (4.3) has a small support contained in $|x|_{\max} \leq \alpha/2$ for some $\alpha > 0$, where $|x|_{\max} := \max_{1\leq i\leq d} |x^i|$. This assumption is easily fulfilled in practice. For instance, for the Gaussian kernel, $K(x) = (2\pi)^{-d/2} \exp(-|x|^2/2)$, which has strictly speaking unbounded support, in practice $K(x)$ is negligible if for some i, $1 \leq i \leq d$, $|x_i| > 6$ 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 kernel estimator is possible. For simplicity we take $t = 0$, $t_1 = T/2$ 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.

Monte Carlo algorithm for the forward-reverse kernel estimator (FRE simulation)

- Simulate N trajectories on the interval $[0, t_1]$, with end points $\{X^{(n)}(t_1) : n = 1, \ldots, N\}$, at a cost of $\mathcal{O}(NLd)$ elementary computations;
- Simulate N reverse trajectories on the interval $[t_1, T]$, 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

$$\begin{aligned}\{X^{(n_k)}(t_1) :\ k = 1, \ldots, l_m\} \quad &:= \quad \{X^{(n)}(t_1) :\ n = 1, \ldots, N\} \\ &\qquad \cap \{x : |x - Y^{(m)}(T)|_{max} \leq \alpha\delta_N\}.\end{aligned}$$

  The size l_m of this intersection is, on average, approximately $N\delta_N^d \times \{$*density of* $X(t_1)$ *at* $Y^{(m)}(T)\}$. It is not difficult to show that this search procedure can be done at cost of order $\mathcal{O}(dN \log N)$;
- Finally, evaluate (4.3) by

$$\frac{1}{N^2\delta_N^d} \sum_{m=1}^{N} \sum_{k=1}^{l_m} K((X^{(n_k)}(t_1) - Y^{(m)}(T))\delta_N^{-1})\mathcal{Y}^{(m)}(T),$$

  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 $d \geq 4$. For $1 \leq d < 4$ 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 FRE algorithm, with $\delta_N = (N/\log N)^{-1/d}$ the FRE simulation requires a total cost of $\mathcal{O}(N \log N)$. Hence, the aggregated costs for achieving $\varepsilon_N \sim 1/\sqrt{N}$ amounts $\mathcal{O}(N \log N)$ which comes down to a complexity $C_\varepsilon^{kern} \sim |\log \varepsilon|/\varepsilon^2$. For $d \geq 4$ we achieve an accuracy rate $\varepsilon_N \sim N^{-\frac{4}{4+d}}$ by taking $\delta_N = N^{-\frac{2}{4+d}}$, again at a cost of $\mathcal{O}(N \log N)$. So the complexity C_ε^{kern} is then of order $\mathcal{O}(|\log \varepsilon|/\varepsilon^{\frac{4+d}{4}})$. For comparison we now consider the classical estimator. It is known that for N trajectories the optimal bandwidth choice is $\delta_N \sim N^{-\frac{1}{4+d}}$, which yields an accuracy of $\varepsilon_N \sim N^{-\frac{2}{4+d}}$. The costs of the classical estimator amounts $\mathcal{O}(N)$ and thus its complexity C_ε^{class} is of order $\mathcal{O}(1/\varepsilon^{\frac{4+d}{2}})$. By comparing the complexities C_ε and C_ε^{class} it is clear that the forward-reverse kernel estimator is superior to the classical Parzen-Rosenblatt kernel estimator for any d.

Complexity of the projection estimator

From its construction in Section 7 it is clear that the evaluation of the projection estimator (4.6) requires 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 $C^{proj}(\varepsilon)$ of order $\mathcal{O}(\log^{d/(4\beta)} |\varepsilon|/\varepsilon^{2+d/(2\beta)})$. If, moreover, the densities $p(t, x, t_1, x')$ and $p(t_1, x', T, y)$ are analytical and the basis functions are sufficiently regular then, (see Remark 7.1) we get root-N accuracy by taking $L_N = \log N$ and so we obtain a complexity of order $C^{proj}(\varepsilon) = |\log \varepsilon|/\varepsilon^2$ for any d. Obviously, compared to the classical estimator, the projection estimator has in any case a better order of complexity when there exists some $\beta > 1$ with β satisfying condition (7.8).

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 $p = 0.518831$, see Figure 1.

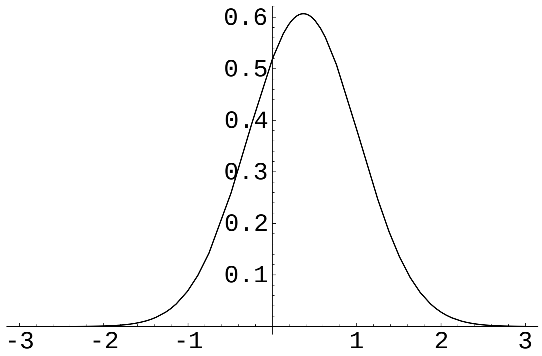

Figure 1: $y \to p(t,x,T,y)$ for $t=0$, $x=1$, $T=1$.

Let us aim to approximate the "true" value $p = 0.518831$ with both the forward-reverse estimator (FRE for short) and the classical forward estimator (FE for short). Throughout this experiment we choose $t_1 = 0.5$ and $M = N$ for the FRE and the FE is simply obtained by taking $t_1 = 1$. For the bandwidth we take $\delta_N^{FE} = N^{-1/5}$ and $\delta_N^{FRE} = N^{-1}$, yielding variances $\sigma_{FE}^2 \approx C_1 N^{-4/5}$ and $\sigma_{FRE}^2 \approx C_2 N^{-1}$, respectively. It is clear that σ_{FE} may be estimated directly from the density estimation since the classical estimator is proportional to a sum of N independent random variables. As the forward-reverse estimator 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, below denoted 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 $\kappa := \sigma_{FRE}/\sigma^*$, see Table 1. The SDEs are simulated by the Euler scheme with time step $\Delta t = 0.01$.

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 1 : 50 FRE simulations

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 forward-reverse estimator may be approximated by $2\kappa\sigma^*$.

In this article we did not address the time discretization error due to the numerical scheme used for the simulation of the SDEs. In fact, this is conceptually the same as 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 which yields 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, for instance, the Talay Tubaro method [13]. The application of this method comes down to Richardson extrapolation of the results obtained by the Euler method for time step $2\Delta t$ and Δt, respectively. However, we have to take into account that the deviation of this extrapolation, and so the Monte Carlo error, is $\sqrt{5}$ times higher. In the experiments below we compare the forward-reverse estimator 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-Tubaro with time discretization steps $\Delta t = 0.02$ and $\Delta t = 0.01$.

N	FRE	$2\sigma_{FRE}$	$\sigma_{FRE}^2 N$	(sec.)	FE	$2\sigma_{FE}$	$\sigma_{FE}^2 N^{4/5}$	(sec.)
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

Table 2 : true $p = 0.518831$.

From Table 2 it is obvious that for larger N the forward-reverse estimator gives a higher Monte Carlo error than the pure forward estimator while the computational effort involved for 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

[1] V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations I: convergence rate of the distribution function. Probab. Theory Related Fields, 104(1996), pp. 43-60.

[2] V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations I: convergence rate of the density. Probab. Monte Carlo Methods Appl., 2(1996), pp. 93-128.

[3] M. Bossy and D. Talay. A stochastic particle method for the McKean-Vlasov and Burgers equations. Math. Comp., 66(1997), (217), 157-192.

[4] L. Devroye and L. Györfi. Nonparametric Density Estimation: The L_1 View. Wiley, 1985.

[5] E.B. Dynkin. Markov Processes. Springer, 1965 (engl. transl. from Russian 1963).

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

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

[8] O. Kurbanmuradov, Ü. Rannik, K. Sabelfeld, T. Vesala. Direct and adjoint Monte Carlo for the footprint problem. Monte Carlo Methods Appl., 5(1999), no. 2, pp. 85-111.

[9] O. Kurbanmuradov, Ü. Rannik, K. Sabelfeld, T. Vesala. Evaluation of mean concentration and fluxes in turbulent flows by Lagrangian stochastic models. Mathematics and Computers in Simulation, 54(2001), pp. 459-476.

[10] G.N. Milstein. Numerical Integration of Stochastic Differential Equations. Kluwer Academic Publishers, 1995 (engl. transl. from Russian 1988).

[11] G.N. Milstein and J.G.M. Schoenmakers. Monte Carlo construction of hedging strategies against multi-asset European claims. To appear in Stochastics and Stochastics Reports.

[12] N.J. Newton. Variance reduction for simulated diffusion. SIAM J. Appl. Math., 54(1994), pp. 1780-1805.

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

[14] D.J. Thomson. Criteria for the selection of stochastic models of particle trajectories in turbulent flows. J. Fluid. Mech., 180(1987), pp. 529-556.

[15] W. Wagner. Monte Carlo evaluation of functionals of solutions of stochastic differential equations. Variance reduction and numerical examples. Stoch. Anal. and Apl., 6(1988), pp. 447-468.