

Transition Densities of Diffusion Processes: Numerical Comparison of Approximation Techniques

BJARKE JENSEN AND ROLF POULSEN

BJARKE JENSEN

is an analyst at SEB Fixed Income Research in Copenhagen, Denmark.
bjarke.jensen@seb.se

ROLF POULSEN

is an assistant professor of statistics and operations research, Institute for Mathematical Sciences, at the University of Copenhagen in Denmark.
rolf@math.ku.dk

This article compares techniques for approximating the transition densities of a diffusion process. It considers Euler approximations, simulation-based methods, binomial approximations, a numerical solution of Fokker-Plank partial differential equations, and Hermite expansions. From a speed/accuracy trade-off perspective of popular models with known densities, this list is ordered from least preferred to best method.

Diffusion processes are characterized by their transition densities, but they are generally not able to be expressed in a closed-form way. Diffusions are widely used in financial modeling, and there are two main reasons to try to determine transition densities: 1) to achieve likelihood-based statistical inference (the Markov property means that the likelihood function based on discrete observations is just the product of one-period transitions); and 2) to price simple contingent claims (the no-arbitrage price is the integral of the discounted payoff with respect to the risk-neutral density).

We do know many features of both the probabilistic and analytical nature of the densities. These features suggest ways to approximate the densities.

We provide a detailed comparison of some of the methods most commonly used. We describe five approximation methods for the transition densities of a diffusion process: the Euler approximation, simulation methods, binomial approximations, numerical solutions of a partial differential equation related to the tran-

sition density, and finally a Hermite expansion. The first four methods are all more or less established in the financial and statistical literature, while the last method has only recently been proposed by Aït-Sahalia [2002] and [1999].

We compare these with standard models in which the transition densities are known: the Vasicek [1977] model, the Cox, Ingersoll, and Ross [1985] model, and the Black and Scholes [1973] model. For financially reasonable parameter values, the least preferred approximation technique in terms of speed/accuracy trade-offs is the Euler approximation, and the best is a Hermite expansion.

We consider a one-dimensional time-homogeneous diffusion process X with dynamics governed by the stochastic differential equation (SDE):

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x, \quad (1)$$

where W is a Brownian motion on some filtered probability space, and $\mu, \sigma: \mathbb{R} \mapsto \mathbb{R}$ are functions such that Equation (1) has a solution with smooth transition densities.¹

It is these transition densities that we want to find. That is, we are looking for $\phi: \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$ (referred to as the *transition density*) such that

$$y \mapsto \phi(\Delta, x, y)$$

is the density of X_{Δ} , given $X_0 = x$.

I. APPROXIMATION TECHNIQUES

Euler Approximations

If Δ is small, we can think of dX_t as $X_\Delta - x$, approximate the X_t on the right-hand side of Equation (1) with x , and treat dW_t as a $N(0, \Delta)$ variable. This leads to the approximation

$$\phi^A(\Delta, x, y) = n(y; x + \mu(x)\Delta, \sigma^2(x)\Delta),$$

where $n(\cdot; \alpha, \beta)$ is the density of the normal distribution with mean α and variance β . This is called the Euler approximation. It is effective when Δ is small; see Kloeden, Platen, and Schurz [1994] for precise statements about the types and orders of convergence.

When we need a density approximation, however, Δ is typically not a quantity the modeler can control. In many situations, we may have a clear feeling that Δ is small, but the error is hard to approximate in a constructive manner.

If the conditional first and second moments, say, m and v , of X are known, then we can use

$$\phi^A(\Delta, x, y) = n(y; m(\Delta, x), v(\Delta, x)),$$

which we call a *correct moment approximation*. This approximation to the density is also accurate only when Δ is small, but at least we have removed one source of approximation error, and that has advantages. For instance, using the approximation for statistical inference leads to martingale estimation functions and thus produces consistent estimates under weak conditions (see Bibby and Sørensen [1995]).

Simulation

When a simulation approach is used to approximate the density for Equation (1), several things have to be considered. The first task is generating random numbers. We consider two generators: pseudo-random and quasi-random.

For a *pseudo-random generator*, we use a linear congruential formula; see Press et al. [1992]. This creates numbers that behave as if they were independent and $U(0, 1)$ -distributed.

We also consider *quasi-random numbers* or *low-discrepancy sequences*, which have recently received some attention. A one-dimensional quasi-random sequence

is a sequence of numbers in the interval $[0, 1]$, where numbers have been carefully chosen to “fill out the space” more rapidly than just picking independent $U(0, 1)$ variables. In this case, the numbers are not random—but very much the opposite, one could say. It is not hard to imagine that this “space-filling property” is useful for Monte Carlo integration, which is what we are doing abstractly in simulations using the law of large numbers. For more details on quasi-random numbers and references to the underlying mathematical literature, see Galanti and Jung [1997] and Joy, Boyle, and Tan [1996].

Quasi-random points can also be defined and created in \mathbb{R}^M , where they are points that fill out the unit cube fast, but it is somewhat more difficult to create M -dimensional quasi-numbers because the coordinates cannot be chosen “independently.” In other words, to get N quasi-points in the \mathbb{R}^M -unit cube, we cannot just take $N \times M$ points from a one-dimensional quasi-sequence, and then rearrange them as N vectors of length M (as we can with pseudo-numbers). We have to be careful not to scramble the ordering of the quasi-numbers; see Joy, Boyle, and Tan [1996] for more details. We use the so-called Sobol numbers, which we get from an extension (to dimensions higher than 6) of the algorithm in Chapter 7 in Press et al. [1992].²

With a uniform sequence, we would like to simulate outcomes of X_Δ by applying the inverse distribution function. Our main point, however, is that we do not know this function; we do not even know the density. The solution is to discretize Equation (1). There are different ways (often called *schemes*) to do this. An excellent survey of the topic can be found in Kloeden, Platen, and Schurz [1994].

We consider two schemes: the Euler and the Milstein schemes.³ They are given by

$$X_{(j+1)\delta} - X_{j\delta} = \delta\mu(X_{j\delta}) + \sqrt{\delta}\sigma(\cdot)n_{j+1}, \quad (2)$$

$$\begin{aligned} X_{(j+1)\delta} - X_{j\delta} = & \delta \left(\mu(X_{j\delta}) - \frac{1}{2}\sigma(\cdot)\sigma'(\cdot) \right) + \sqrt{\delta}\sigma(\cdot)n_{j+1} + \frac{\delta}{2}\sigma(\cdot)\sigma'(\cdot)n_{j+1}^2 \\ & + \delta^{3/2} \left(\frac{1}{2}\mu(\cdot)\sigma'(\cdot) + \frac{1}{2}\mu'(\cdot)\sigma(\cdot) + \frac{1}{4}\sigma^2(\cdot)\sigma''(\cdot) \right) n_{j+1} \\ & + \delta^2 \left(\frac{1}{2}\mu(\cdot)\mu'(\cdot) + \frac{1}{4}\mu''(\cdot)\sigma^2(\cdot) \right) \end{aligned} \quad (3)$$

where a prime mark denotes differentiation; a bullet indicates that the function should be evaluated at $X_{j\delta}$ (as all right-hand side functions are); $X_0 = x$ is given; $\delta = \Delta/M$ is the time step (chosen by us) between two discretizations; M is a fixed number of discretization times; and n_1, n_2, \dots, n_M are

standard normal variables.⁴ The Milstein scheme is a refinement of the Euler scheme leading to higher accuracy (weak-second-order as opposed to weak first-order convergence).

Generating M standard normal variables therefore leads to one realization of X_Δ by using either the Euler or the Milstein scheme, and $N \times M$ variables are needed to produce N numbers used to estimate the density of Equation (1). Whichever generator and simulation scheme we use, we end up with N realizations of X_Δ .

As shown in Duffie and Glynn [1995] and in Duffie [1996], there is a computationally optimal relationship (depending on the order of accuracy of the scheme) between the number of simulations (N) and the number of time steps (M) with a fixed amount of computer time. It is asymptotically optimal to quadruple the number of simulations with each doubling of time steps for the Euler scheme. For the Milstein scheme, one should make 16 times as many simulations when the number of time steps is doubled. These results also explain why there is not much gained by considering higher-order schemes; it's the simulation error, not the discretization error, that matters.

The N realizations of X_Δ can be used to approximate the density for (1) or to approximate the price of a financial security. To estimate a transition density, one basically has to make a histogram on the real axis \mathbb{R} or the half-axis $[0, \infty)$, whichever is the supporting region for the underlying process X . Suppose that the relevant axis is divided into intervals, $[y_i, y_{i+1})$. The simulation of one realization of X_Δ is in one of the intervals $[y_i, y_{i+1})$, and after N simulations we count how many realizations actually end up in the interval $[y_i, y_{i+1})$. Dividing each of these numbers by N , we have the histogram that approximates the transition density.

One could also use kernel smoothing, known from non-parametric statistical analysis, although a histogram is actually a kind of kernel smoothing, with the choice of kernel being the rectangular one. Other kernels could be chosen, such as the Gaussian density function. For our purposes, there is not much to gain using other kernels. This statement is documented in Silverman:

There is very little to choose between the various kernels on the basis of mean integrated square error. It is perfectly legitimate, and indeed desirable, to base the choice of kernel on other considerations, for example the degree of differentiability required or the computational effort involved [1986, p. 43].

Since differentiability of the density is not an issue here, and since the rectangular kernel is the least computationally time-consuming kernel, a histogram perfectly satisfies the demands of our numerical analysis of transition densities.

For the N realizations of X_Δ generated by a pseudo-generator, the error is of order $1/\sqrt{N}$ in the central limit theorem sense. The central limit theorem does *not* apply to the quasi-generator. Instead, it can be shown that the difference between true quantities and the sample counterparts is of the order $(\ln N)^M/N$, which is smaller than the error induced using pseudo-numbers. As we will show, quasi-numbers are not overwhelmingly more efficient than pseudo-numbers, especially when computer time is considered. A further problem with quasi-numbers is that we have no simple way to approximate or estimate the extent of the error in a typical finite sample.

A final comment regarding other techniques for improving simulation is in order. First, the control variate techniques must be mentioned. These methods depend on a relationship between a known (analytical) density and the density for Equation (1) found by simulation.

Suppose we use the same random numbers to create approximations to both the density for (1) and the known density. If the control problem looks like the original one, then there is good reason to believe that the errors (one known, one unknown) are the same. This can be used to correct our estimate of the unknown density. What makes a good control variate is highly problem-dependent, and we therefore do not implement such methods in this article (although using the standard normal density might improve the pseudo-random estimate).

Another way to reduce the variance is the antithetic technique, where the basic idea is simply that a realization of a symmetric random variable Z is paired with the identical distributed random variable $-Z$. This is used to generate two realizations of X_Δ , the first using the realizations n_1, n_2, \dots, n_M , and the second $-n_1, -n_2, \dots, -n_M$. Since this method is generic, we use it in both the pseudo- and the quasi-random experiments.

Binomial Models

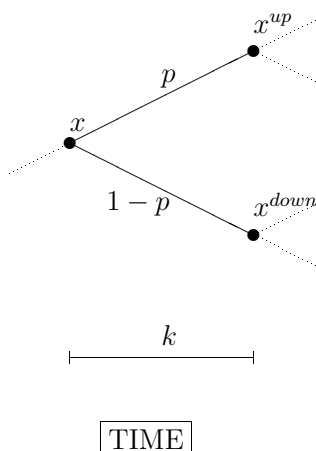
Binomial models are classic all-purpose tools in finance, so we all know what they look like. Exhibit 1 shows zooming in on a particular node.

When we want a binomial model to approximate a diffusion model, which we assume is given by (1), we have to let the time step-size k tend to zero; let the number of

EXHIBIT 1

Binomial Model Node

SPACE



steps tend to infinity (equally fast of course); and choose space step-sizes and probabilities appropriately. To ensure convergence of a binomial model, two conditions must be met:⁵

1. The local first- and second-moments of the binomial process must converge to the drift and volatility of the diffusion process (and do so in a uniform way).
2. The jump sizes in the binomial model must tend to 0 (also appropriately uniformly).

This still leaves us considerable freedom in the choice of probabilities and up and down moves. One way to construct a binomial model is first to try with:⁶

$$\begin{aligned}x^{up} &= x + \sqrt{k}\sigma(x), \\x^{down} &= x - \sqrt{k}\sigma(x), \\p &= \frac{1}{2} + \frac{1}{2}\sqrt{k}\frac{\mu(x)}{\sigma(x)}.\end{aligned}$$

This essentially takes care of conditions 1) and 2); minor (but tedious) adjustments may be necessary to ensure uniformity (a problem that arises if the σ function is not bounded away from 0).

This plan has one serious flaw: Unless σ is constant, it will not lead to a recombining model (a lattice), so computational burdens will grow exponentially.⁷

Suppose we can find a function f such that

$$f'(z) = \frac{1}{\sigma(z)}$$

in the set where $\sigma(z) > 0$. Then the Ito formula tells us that the process $(Y_t) = [f(X_t)]$ has unit volatility.

Our “first try” leads to a lattice model for Y . Since f is strictly increasing (forgetting for a moment about the 0s of σ), it has an inverse function that we can use to transform the Y lattice into an X lattice. Convergence in distribution is preserved under continuous transforms, and we are done.

This is the basic idea; some caution is needed to deal with the singularities of $1/\sigma$. This is seen in Nelson and Ramaswamy [1989], who provide both a general treatment and specific formulas for commonly used models. This idea of creating a diffusion with unit volatility is also used in the Hermite expansion.

The final thing we need to do is to use the binomial model to estimate the transition density of the diffusion. For a given initial condition x and a Δ value, we estimate $\phi(x, \Delta, y)$ at y values corresponding to values at terminal nodes (at Δ) of the (X) lattice as

$$\frac{\text{probability of ending at } y}{1/2 * \text{distance to the point below } y + 1/2 * \text{distance to point above } y}$$

The cumulative probabilities are easily found from the local ones when we work through the lattice. Explicit formulas for the specific diffusion models considered can be found in Nelson and Ramaswamy [1989].

Partial Differential Equation Techniques

Under regularity conditions, ϕ solves the parabolic partial differential equation (PDE) called the Fokker-Plank or the Kolmogorov forward equation (see Karatzas and Shreve [1992]):⁸

$$\begin{aligned}\frac{\partial}{\partial t}\phi(t, x, y) &= -\frac{\partial}{\partial y}(\mu(y)\phi(t, x, y)) + \\&\quad \frac{1}{2}\frac{\partial^2}{\partial y^2}(\sigma^2(y)\phi(t, x, y)),\end{aligned}\quad (4)$$

with initial condition $\phi(0, x, y) = \delta(y - x)$, where $\delta(\cdot)$ is the Dirac δ function. If we are careful, quite useful approximations to the solution of this PDE can be

obtained using finite-difference methods, such as the Crank-Nicolson method (for details, see Strikwerda [1989] or Duffie [1996]).

In finite-difference methods, the time and space variables are discretized (say, into step-sizes k and h , respectively), and approximate solutions are found by solving sequences of sparse linear systems (typically tridiagonal). The Crank-Nicolson method is locally second-order accurate in both h and k (only $O(h^3, k^3)$ terms remain in a Taylor expansion) and unconditionally stable for parabolic PDEs. This lets us use the Lax-Richtmeyer equivalence theorem (Strikwerda [1989, Theorem 10.5.1]) to conclude that the numerical solution converges.

This theorem, however, does not give any statement about the order of convergence. We can hope that the local properties (second-order accuracy) are inherited globally, but this has to be checked numerically. To do this we use the error analysis technique suggested in Østerby [1998], which is based on the same arguments as the well-known Richardson extrapolation technique that we then apply (see Poulsen [1999, Section 3.2] for further details).

We use as the initial condition at some point y_i in the discretized world a normal-density approximation:

$$n(y_i; x + \mu(x)k, \sigma^2(x)k), \quad (5)$$

where x is the starting point of the diffusion. By one definition of the Dirac δ function, this converges to the true initial condition as $k \rightarrow 0$. As pointed out in Strikwerda [1989, p. 121], the Crank-Nicolson method tends to behave badly for non-smooth initial data (the Dirac δ function is not a smooth function; it isn't even strictly a function), but with the choice above things work nicely.⁹

Clearly, if the order of accuracy is the same in the time dimension as in the space dimension, it is computationally efficient to keep the step sizes proportional, i.e., let them tend to 0 at the same rate. The error/order analysis mentioned above can be used to find a good constant of proportionality.

Hermite Expansion

The idea of expanding an unknown density function in terms of its moments by using Hermite polynomials goes back to Cramér [1925]. He also shows, however, that the density has to be almost normal for the Hermite expansion to converge. Most transition densities in finance are not.

Ways of expanding unknown and far-from-normal densities in terms of moments/cumulants have been developed. The Edgeworth expansion is one example used in finance to approximate the density of a sum of lognormal variables (needed for pricing Asian options).

Rather than working on the non-normal densities directly, Aït-Sahalia's [1999] idea is to transform the density into something that is close to normal and then expand on that. The advantage is that in a diffusion setting the appropriate transform and its inverse can be found using only the drift and volatility functions; no further knowledge of the unknown density is needed. In this way, the method is parsimonious; a successful Edgeworth expansion of the untransformed density would require a more detailed investigation of *how* it is non-normal (which would probably amount to the same calculations as Aït-Sahalia performs).

We briefly review the transformations, assumptions, and main results for Hermite expansions. For more in-depth analysis, see Aït-Sahalia [1999 and 2002].

We first transform X into $Y = f(X)$, where Y satisfies

$$dY_t = \mu_Y dt + dW(t),$$

The Ito formula tells us that $f' = 1/\sigma$ and $\mu_Y(y) = \mu f' - \sigma f''/2 = \mu(f^{-1}(y))/\sigma[f^{-1}(y)] - \sigma'[f^{-1}(y)]/2$. We then further center and scale into

$$Z = \sqrt{\Delta}(Y_\Delta - y_0).$$

Now define the Hermite polynomials as:¹⁰

$$H_j(z) := e^{\frac{z^2}{2}} \frac{d^j}{dz^j} \left[e^{-\frac{z^2}{2}} \right], \quad j \geq 0.$$

Define the J -th truncated density of Z (which will serve as an approximation of the true density ϕ_Z) as

$$\phi_Z^{(J)}(\Delta, z_0, z) := n(z) \sum_{j=0}^J \eta_j(\Delta, z_0) H_j(z) \quad (6)$$

where $n(z)$ is the standard normal density, and

$$\eta_j(\Delta, z_0) := \frac{1}{j!} \int_{-\infty}^{+\infty} H_j(u) \phi_Z(\Delta, z_0, u) du.$$

If the η_j functions can be determined, the right-hand side of (6) can be calculated, and we have an approximation of ϕ_Z . By the Jacobian transformation result, the J -th truncated densities of Y and X can be found as

$$\phi_Y^{(J)}(\Delta, y_0, y) = \Delta^{-\frac{1}{2}} \phi_Z^{(J)}\left(\Delta, y_0, \Delta^{-\frac{1}{2}}(y - y_0)\right) \quad (7)$$

$$\phi_X^{(J)}(\Delta, x_0, x) = \frac{\phi_Y^{(J)}(\Delta, f(x_0), f(x))}{\sigma(x)}, \quad (8)$$

so we have a direct approximation of the X density. This is important in a likelihood context, because we can specify (and optimize) the (approximate) likelihood function for observed data directly; we do not have to apply a parameter-dependent transform to the data (which would force us to “carry a Jacobian around”).

These truncated densities are indeed good approximations, because Ait-Sahalia [2000] shows that under general conditions (stationarity for instance):

$$\phi_X^{(J)}(\Delta, x_0, y) \rightarrow \phi(\Delta, x_0, y), \quad \text{for } J \rightarrow \infty,$$

for all x_0, y in the domain of the diffusion.

Suppose now that the f transformation and its inverse have been found, and some J sufficiently large for our liking has been fixed. We then need the coefficients $\eta_j(\Delta, y_0)$, $j = 0, \dots, J$, to compute $\phi_Z^{(j)}(\Delta, y_0, z)$ —and seemingly we’re back to square one, because these depend on the true and unknown density ϕ_Z . But this too can be solved by some further approximations/expansions.

Explicit calculations yield:

$$\eta_j(\Delta, y_0) = \frac{1}{j!} E\left(H_j\left(\Delta^{-\frac{1}{2}}(Y_\Delta - y_0)\right) | Y_0 = y_0\right).$$

This expectation can be evaluated using a Taylor approximation. Ait-Sahalia [2000] shows that for a smooth function g and an integer n , we have

$$E[g(Y_\Delta) | Y_0 = y_0] = \sum_{i=0}^n (\mathcal{A}^i g)(y_0) \frac{\Delta^i}{i!} + \text{remainder}$$

where \mathcal{A} is the infinitesimal operator of the diffusion Y given by

$$(\mathcal{A}g)(y_0) = \mu_Y(y_0)g'(y_0) + \frac{1}{2}g''(y_0)$$

(recall that Y has unit volatility), and $\mathcal{A}^i g$ means \mathcal{A} applied i times to g .¹¹

As usual in Taylor expansions, the remainder includes a $1/n!$ -factor, and when g is regular (say, has at most exponential growth) this sends it to zero as n grows. The idea is to choose some n and then apply this result, with each of the Hermite polynomials H_0, \dots, H_J (remember J has been fixed already) playing the role of g , thus obtaining approximations to η_0, \dots, η_J .

Ait-Sahalia [2000] suggests that one decides first on the truncation point J , and then chooses n so that the Taylor expansion has terms of at most order $\Delta^{J/2}$. Following this suggestion, we fix $J = 6$, and with the help of the computer program Mathematica we find the expressions for the η_j ($\mu_Y^{(j)}$ means the j -th derivative):

$$\begin{aligned}
 \eta_0 &= 1, \\
 \eta_1 &= -\mu_Y \sqrt{\Delta} - \left(\frac{1}{2} \mu_Y \mu'_Y + \frac{1}{4} \mu''_Y \right) (\sqrt{\Delta})^3 \\
 &\quad - \left(\frac{1}{6} \mu_Y (\mu'_Y)^2 + \frac{1}{6} \mu_Y^2 \mu''_Y + \frac{1}{6} \mu_Y \mu_Y^{(3)} + \frac{1}{4} \mu'_Y \mu''_Y + \frac{1}{24} \mu_Y^{(4)} \right) (\sqrt{\Delta})^5, \\
 \eta_2 &= \frac{1}{96} \Delta \left(\begin{aligned} &48 \mu_Y^2 + 48 \mu'_Y + 48 \Delta \mu_Y^2 \mu'_Y + 56 \Delta \mu_Y \mu''_Y + 32 \Delta (\mu'_Y)^2 + 16 \Delta \mu_Y^{(3)} \\ &+ 16 \Delta^2 \mu_Y^3 \mu''_Y + 3 \Delta^2 \mu_Y^{(5)} + 21 \Delta^2 (\mu''_Y)^2 + 16 \Delta^2 (\mu'_Y)^3 \\ &+ 28 \Delta^2 \mu_Y^2 \mu_Y^{(3)} + 16 \Delta^2 \mu_Y \mu_Y^{(4)} \\ &+ 32 \Delta^2 \mu'_Y \mu_Y^{(3)} + 88 \Delta^2 \mu_Y \mu''_Y \mu'_Y + 28 \Delta^2 \mu_Y^2 (\mu'_Y)^2 \end{aligned} \right), \\
 \eta_3 &= \frac{1}{6} \left(\begin{aligned} &-3 (\sqrt{\Delta})^3 \mu_Y \mu'_Y - (\sqrt{\Delta})^3 \mu''_Y - \frac{7}{2} (\sqrt{\Delta})^5 \mu_Y (\mu'_Y)^2 \\ &-\frac{11}{4} (\sqrt{\Delta})^5 \mu_Y^2 \mu''_Y - (\sqrt{\Delta})^3 \mu_Y^3 - \frac{7}{4} (\sqrt{\Delta})^5 \mu_Y \mu_Y^{(3)} \\ &-3 (\sqrt{\Delta})^5 \mu'_Y \mu''_Y - \frac{3}{8} (\sqrt{\Delta})^5 \mu_Y^{(4)} - \frac{3}{2} (\sqrt{\Delta})^5 \mu_Y^3 \mu'_Y \end{aligned} \right), \\
 \eta_4 &= \frac{1}{240} \Delta^2 \left(\begin{aligned} &60 \mu_Y^2 \mu'_Y + 40 \mu_Y \mu''_Y + 30 (\mu'_Y)^2 + 10 \mu_Y^{(3)} + 50 \mu_Y^3 \mu''_Y \Delta + 4 \mu_Y^{(5)} \Delta \\ &+ 34 (\mu''_Y)^2 \Delta + 40 (\mu'_Y)^3 \Delta + 50 \mu_Y^2 \mu_Y^{(3)} \Delta + 23 \mu_Y \mu_Y^{(5)} \Delta + 52 \mu'_Y \mu_Y^{(3)} \Delta \\ &+ 180 \mu_Y \mu_Y'' \mu'_Y \Delta + 10 \mu_Y^4 + 100 \mu_Y^2 (\mu'_Y)^2 \Delta + 20 \Delta \mu_Y^4 \mu'_Y \end{aligned} \right), \\
 \eta_5 &= -\frac{1}{120} (\sqrt{\Delta})^5 \left(5 \mu_Y \mu_Y^{(3)} + 10 \mu_Y^3 \mu'_Y + 15 \mu_Y (\mu'_Y)^2 + 10 \mu'_Y \mu''_Y + 10 \mu_Y^2 \mu''_Y + \mu_Y^5 + \mu_Y^{(4)} \right), \\
 \eta_6 &= \frac{1}{720} \Delta^3 \left(\begin{aligned} &10 (\mu''_Y)^2 + \mu_Y^{(5)} + 15 (\mu'_Y)^3 + 45 \mu_Y^2 (\mu'_Y)^2 + 6 \mu_Y \mu_Y^{(4)} + 15 \mu'_Y \mu_Y^{(3)} + 20 \mu_Y^3 \mu''_Y \\ &+ 15 \mu_Y^4 \mu'_Y + 15 \mu_Y^2 \mu_Y^{(3)} + 60 \mu_Y \mu''_Y \mu'_Y + \mu_Y^6 \end{aligned} \right).
 \end{aligned}$$

Using (6) we get approximations for $\phi_Z^{(j)}(\Delta, x_0, x)$, and with (7) and (8) we get approximations for $\phi_X^{(j)}(\Delta, x_0, x)$, all for $j = 0, \dots, 6$. These are long and imposing expressions, but still closed-form approximations that can be made as accurate as one wants.

II. MODELS WITH KNOWN TRANSITION DENSITIES

There are three models that are well-known to the financial community and have closed-form expressions for the transition density.

Vasicek Model

We say Vasicek model in tribute to Vasicek [1977]; people outside the financial community would call it an Ornstein-Uhlenbeck process. The dynamics of the SDE are

$$dX_t = \kappa(\theta - X_t)dt + \sigma_V dW_t,$$

for $\kappa, \sigma_V \in \mathbb{R}_+$ and $\theta \in \mathbb{R}$. In this case $X_\Delta | X_0 = x$ is normally distributed; specifically, the transition density is

$$\phi^V(\Delta, x, y) = n(y; m(\Delta, x), v(\Delta, x)),$$

where the conditional moments are

$$\begin{aligned} m(t, x) &= E(X_t | X_0 = x) = e^{-\kappa t} x + \theta(1 - e^{-\kappa t}), \\ v(t, x) &= \text{Var}(X_t | X_0 = x) = \frac{\sigma_V^2(1 - e^{-2\kappa t})}{2\kappa}. \end{aligned}$$

The relevant f transformation to get unit volatility is $x \rightarrow x/\sigma_V$, and it is easy to see that $\mu_Y = (\kappa\theta) = \sigma_V - \kappa\gamma$.

Cox, Ingersoll, and Ross Model

For the Cox, Ingersoll, and Ross (CIR) [1985] model, the dynamics of the SDE are:

$$dX_t = \kappa(\theta - X_t)dt + \sigma_{CIR}\sqrt{X_t}dW_t,$$

for $\kappa, \theta, \sigma_{CIR} \in \mathbb{R}_+$ such that $2\kappa\theta > \sigma_{CIR}^2$. In this case $X_\Delta | X_0 = x$ follows a non-central χ^2 distribution; specifically, the transition density has support on \mathbb{R}_+ and is

$$\phi^{CIR}(\Delta, x, y) = c \exp(-u - v) \left(\frac{v}{u}\right)^{q/2} I_q(2\sqrt{uv}) \quad \text{for } x, y \in \mathbb{R}_+,$$

where

$$\begin{aligned} c &= \frac{2\kappa}{\sigma_{CIR}^2(1 - \exp(-\kappa\Delta))}, & q &= \frac{2\kappa\theta}{\sigma_{CIR}^2} - 1, \\ u &= cx \exp(-\kappa\Delta), & v &= cy, \end{aligned}$$

and I_q is the modified Bessel function of the first kind of order q .

The conditional mean is given by the same expression as for the Vasicek model, while

$$v(t, x) = x \frac{\sigma_{CIR}^2(e^{-\kappa t} - e^{-2\kappa t})}{\kappa} + \theta \frac{\sigma_{CIR}^2(1 - e^{-\kappa t})^2}{\kappa}.$$

The relevant f transformation is $x \mapsto (2\sqrt{x})/\sigma_{CIR}$, and we find that $\mu_Y(y) = (2\kappa\theta - \sigma_{CIR}^2/2)/(2\sigma_{CIR}^2\gamma) - \kappa\gamma/2$.

Black-Scholes Model

People outside the financial community would call the Black-Scholes [1973] model a geometric Brownian motion with drift. The dynamics of the SDE are:

$$dX_t = rX_t dt + \sigma_{BS}X_t dW_t,$$

for $r, \sigma_{BS} \in \mathbb{R}_+$. In this case $X_\Delta | X_0 = x$ is lognormally distributed; specifically, the transition density is

$$\begin{aligned} \phi^{BS}(\Delta, x, y) &= \frac{1}{\sigma_{BS}y\sqrt{\Delta}\sqrt{2\pi}} \exp\left(-\frac{(\ln y - (\ln x + (r - \sigma_{BS}^2/2)\Delta))^2}{2\sigma_{BS}^2\Delta}\right) \\ &\quad \text{for } x, y \in \mathbb{R}_+, \end{aligned}$$

and

$$\begin{aligned} m(t, x) &= xe^{rt}, \\ v(t, x) &= x^2 e^{2rt} (e^{\sigma_{BS}^2 t} - 1). \end{aligned}$$

The relevant f transformation is $x \rightarrow (\ln x)/\sigma_{BS}$, and $\mu_Y(y) = r/\sigma_{BS} - \sigma_{BS}/2$.

III. NUMERICAL RESULTS

Since the object we are trying to approximate is a whole function, there are many ways to measure the quality of the approximation.¹² We give two error measures, one absolute, one relative. (The y_i is the point where the approximate solution is calculated.)

- Maximal absolute error:

$$e_1 = \sup_i \{|\phi^A(y_i) - \phi(y_i)|\}.$$

Evidently, this is a reasonable way to measure the error, but generally it does not make sense to compare e_1 across models.

- Average relative error (quoted in parts per million in exhibits):

$$e_2 = \sum_i |\phi^A(y_i) - \phi(y_i)| \Delta y \\ \approx \int_{\mathbf{R}} |\phi^A(y) - \phi(y)| dy = \int_{\{y|\phi(y)>0\}} \frac{|\phi^A(y) - \phi(y)|}{\phi(y)} \phi(y) dy.$$

The relative error tells us how many significant digits of the solution we can trust, and therefore the errors can be compared across models. One might wonder why we do not measure, say, the maximal relative error. This does not produce very informative numbers, since the functions we are working with are very close to 0 in large areas. But we have a reasonable tool for measuring the importance of the relative errors, namely, the function itself.

It would be possible to use more application-oriented error measures, such as the degree of option mispricing or the extent to which we lose the good statistical properties of the maximum-likelihood estimation by using an approximation.

We use financially realistic parameters shown in Exhibit 2 in our numerical experiments (see Chan et al. [1992], for instance). More technique-specific parameters are indicated in the relevant notes. *Time* is the CPU time (in seconds) on an HP-9000 Unix machine. We present results in tabular form only for the analytical approximations (the simple and the Hermite). For the numerical methods

EXHIBIT 2

Parameters for Numerical Experiments

Model	Parameters			
Vasicek	$\Delta = 1/12$	$x = \theta = 0.08$	$\kappa = 0.24$	$\sigma_V = 0.025$
CIR	$\Delta = 1/12$	$x = \theta = 0.08$	$\kappa = 0.24$	$\sigma_{CIR} = 0.08838$
BS	$\Delta = 1$	$x = 100$	$r = 0.08$	$\sigma_{BS} = 0.25$

EXHIBIT 3

Errors for Analytical Normal Approximations of Densities

Model	Euler		“Correct Moment”	
	e_1	e_2	e_1	e_2
Vasicek	0.5535	9643	0	0
CIR	2.001	35164	1.79	33996
BS	0.0027	188990	0.0031	191891

e_1 absolute maximal errors; e_2 average errors in parts per million.

(simulation, binomial, and Crank-Nicolson), information is presented more efficiently in graph form.¹³

Skeptics would say that we are using too small Δ values to make interesting or reasonable comparisons; for such small time steps all densities are quite close to normal. Exhibit 3 shows results for the simple analytical approximations, and reveals that there is no problem detecting non-normality. This does not mean that quantities found using such approximations, typically estimators or derivative prices, are unreasonable, but it is encouraging for studies of refined density approximations.

EXHIBIT 4

Errors of Approximations of CIR Density for Hermite Expansions by Order

Order	e_1	e_2	Time
One	0.514	8940	0.04
Two	0.0289	525	0.05
Three	0.00592	98	0.06
Four	0.000524	11	0.07
Five	0.0000514	0.91	0.09
Six	4.02×10^{-6}	0.013	0.09

e_1 maximal absolute errors; e_2 average relative errors in parts per million.

In Exhibit 4 we see that very low errors are produced for all the Hermite approximations to the CIR density. (The result is the same for the Vasicek and BS models.) The approximations are not very time-consuming, and with just four terms in the expansion we get errors below what any of the other methods can produce.

The graphs in Exhibit 5 show the results of different simulations. For a small number of time discretization steps (M), the Milstein scheme is superior to the Euler scheme. When M is great, there is not much difference. We see that—as expected—for a large number of simulations (N), the errors e_1 and e_2 for pseudo-number approximations decline at the rate $1/\sqrt{N}$. More sur-

prisingly, this also appears to be the case when quasi-numbers are used, which means that the asymptotic properties of quasi-numbers have not kicked in yet. Similar findings in comparisons of quasi- and pseudo-numbers are reported in Berman [1998].

We find that the number of simulations is more important than the number of time steps, which supports the theorem regarding the efficient trade-off between the number of simulations and the number of time approximations found in Duffie and Glynn [1995]. All other things equal (simulation scheme, M , N , and type of kernel estimation), the errors produced by the quasi-numbers tend to be marginally lower than those from pseudo-numbers (something like 10%), but this advantage is wiped out when we take into account the time needed to produce the numbers (about 20% more).

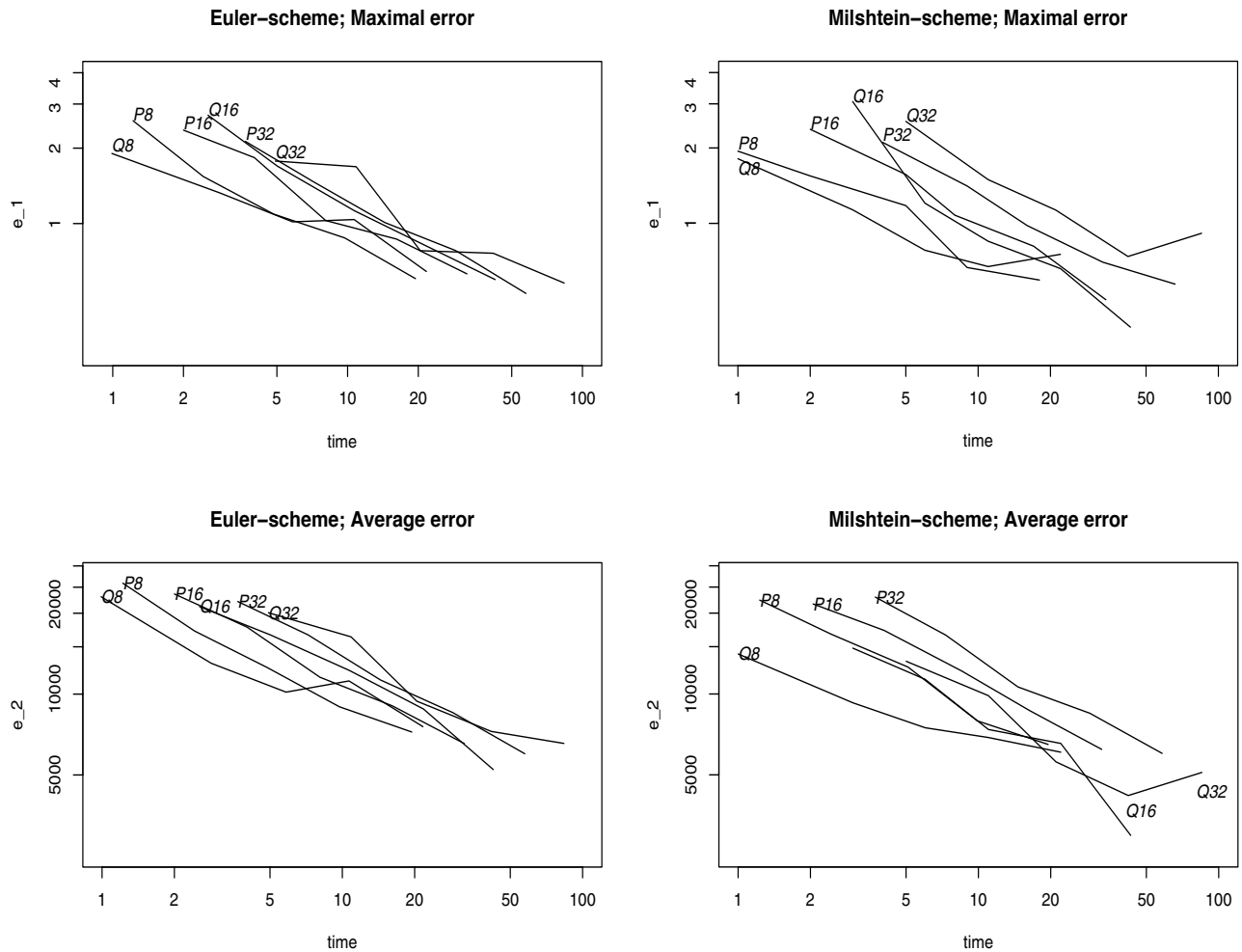
The graphs in Exhibits 6 and 7 show the errors for the binomial and Crank-Nicolson methods with and without extrapolation. As expected, the Crank-Nicolson method performs better than the binomial. There does not appear to be much gain from extrapolation, especially for the Crank-Nicolson method, although looks are a bit deceiving. First, the plots are on a (log, log) scale. Second, the numerical error analysis involved in the extrapolation has been used to determine the optimal ratio between time and space steps. We also see that for very low step sizes the errors tend to flatten out. We attribute this to problems with the computer arithmetic (performing many calculations with very small numbers).¹⁴

The six graphs of the (time, error) trade-off in Exhibits 6 and 7 most strongly convey our message, since they give a clear picture of the ranking of the methods considered. When we take into account the computational effort needed to produce the quasi-numbers, they are no better than pseudo-random numbers. These simulation methods are easily bettered by the binomial method. Binomial approximation is then outperformed by the technique that uses the Crank-Nicolson method to solve the forward PDE.

But neither of those methods can compete with the Hermite expansion. The technique does require some rather tedious algebraic calculations, but they seem to be well worth the effort. We also note a striking similarity of the figures across models and error types. Non-Gaussianity, even non-stationarity (the BS-model), does not affect the picture.

EXHIBIT 5

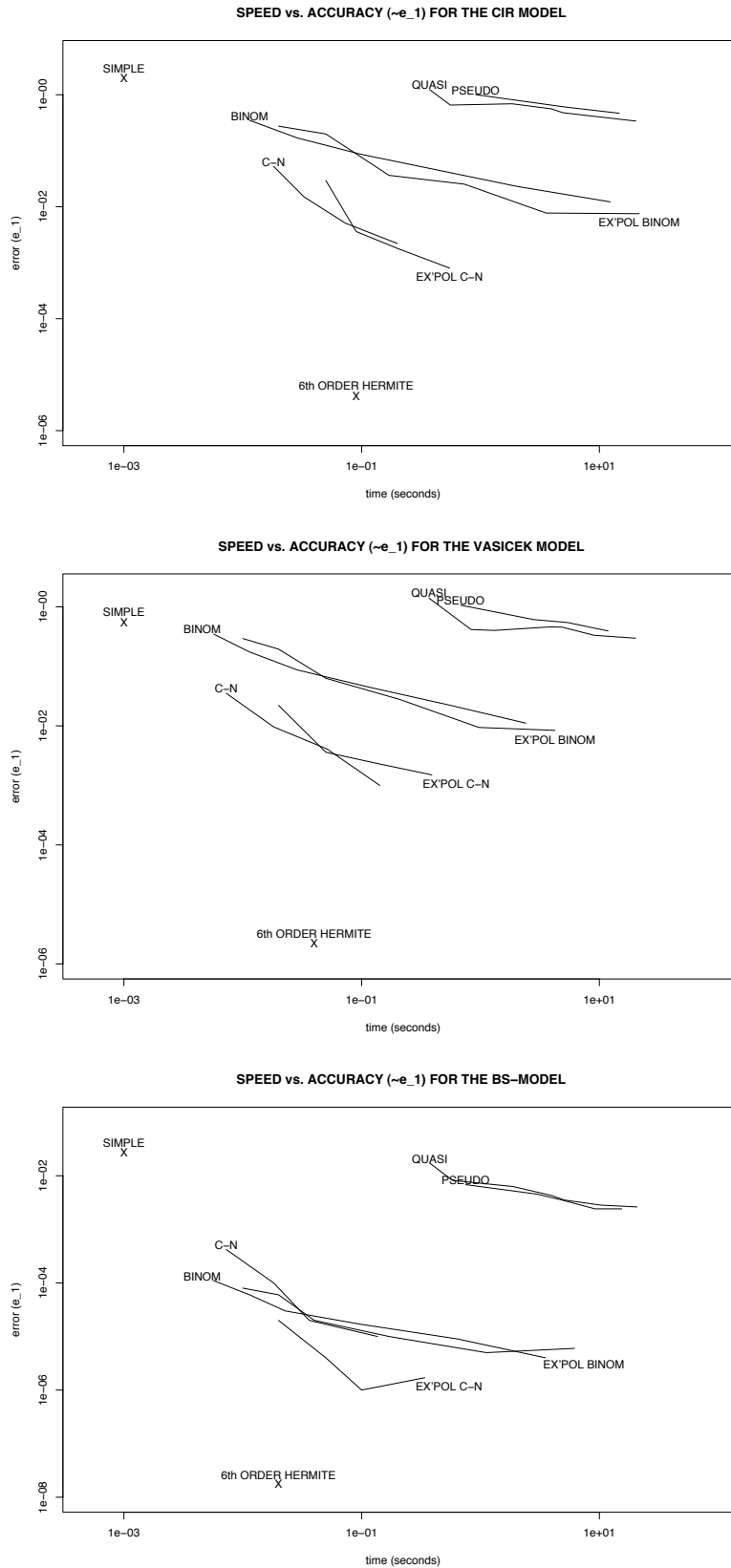
Log/Log Plots of Speed versus Accuracy for Simulation Methods



All graphs are for the CIR Model and 100 interval histograms used in the density estimation. P indicates pseudo-random numbers; Q indicates quasi-random numbers (a Sobol sequence). 8, 16, and 32 indicate number of steps used in time discretization (M). Lines represent increases in time consumption (and reductions in errors) caused by more repetitions (N ranging from 4,000 to 64,000).

EXHIBIT 6

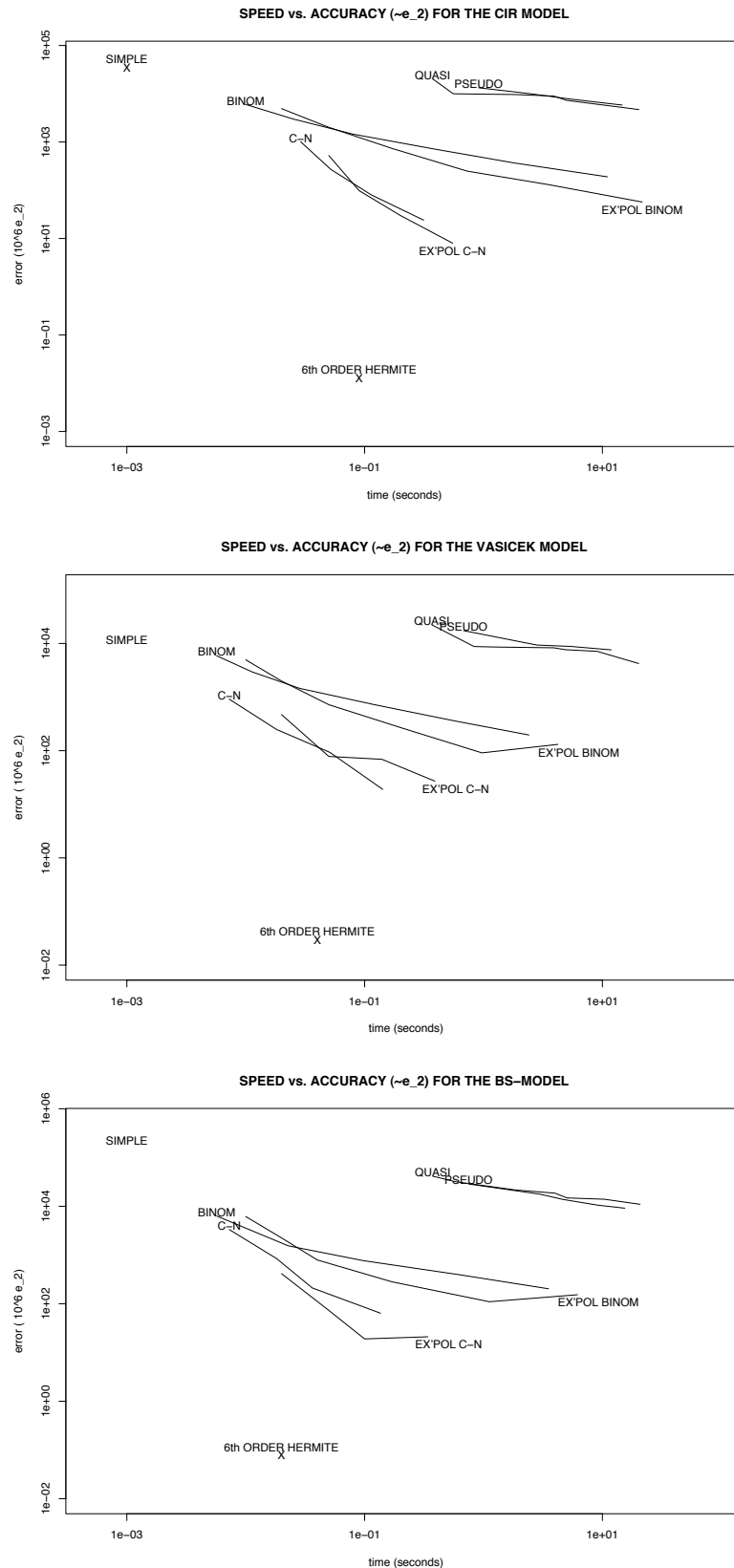
Log/Log Plots of Speed versus Accuracy for Approximation Techniques—Absolute Errors



Crank-Nicolson (C-N) method uses optimal ratio between time and space steps ($M/N \approx 12$, i.e., many fewer time than space steps). Hermite expansion is a 6th-order approximation. Once the technique is implemented, the time consumption for the Hermite expansion is almost constant across orders.

EXHIBIT 7

Log/Log Plots of Speed versus Accuracy—Average Relative Errors



IV. CONCLUSION

A clear ranking of the different approximation techniques emerges from this study. Exhibits 6 and 7 give the result when we compare the two different errors e_1 (maximal absolute) and e_2 (average relative) and the computer time consumed.

Aït-Sahalia's Hermite polynomial expansion is clearly the best method in terms of the speed/accuracy trade-off. The second-best approximation technique is to use the Crank-Nicolson method to solve the PDE for the transition density. The third-best approximation stems from the binomial technique. By far the worst way to approximate the density is the simulation approach described first, whether we use pseudo-random or quasi-random numbers.

An obvious extension of this study would be to apply the approximation techniques to models where we do not know the transition density, since this is of course where we would use the approximations in the first place. All the methods are so general that this is fairly easy to do.

Another topic for future research is multidimensional diffusions. Such extensions are possible, but they are not straightforward, either in theory or in practice. For a two-dimensional diffusion, for instance, different techniques are needed in the PDE approach to achieve a reasonable order of time consumption. In principle, one can use Hermite expansions for two-dimensional diffusions, or apply them to jump processes (processes driven by Levy processes other than Brownian motion), but development along these lines is still in its early stages. Simulations of multidimensional diffusions, however, would not be a problem, so simulations might be the appropriate choice in the case of multidimensional diffusions.

ENDNOTES

The authors thank Yacine Aït-Sahalia, Bent Jesper Christensen, Peter Løchte Jørgensen, David Lando, Kristian Miltersen, Jørgen Aase Nielsen, Jesper Lund Pedersen, Carsten Sørensen, and Michael Sørensen for valuable comments and suggestions; and an anonymous referee who pointed out some mistakes. At the time of this writing, the authors were at the University of Aarhus in Denmark. The authors thank the Centre for Analytical Finance, Aarhus, and the Danish Network in Mathematical Finance for support.

¹Whether a density actually exists depends on the μ and σ functions. Karatzas and Shreve [1992, Section 5.7] give sufficient conditions. How much these can be relaxed and exactly how smooth the density is (possibly as a function of a "statistical" parameter) is a very difficult probabilistic question.

²We also tried the Faure numbers mentioned in Joy, Boyle, and Tan [1996]. The errors behave roughly the same way, but the Faure numbers take longer to create (by a factor of about 7 in our implementations).

³In the literature, the term "Milshtein scheme" often refers to a scheme including only the first line of Equation (3). This scheme has a lower convergence order. Both schemes were proposed by Milshtein [1978]. We are grateful to an anonymous referee for clarifying this.

⁴To transform the uniform numbers into standard normals, we use the Box-Muller-transform for the pseudo-numbers, and Moro's approximation of the inverse cumulative normal distribution for the quasi-numbers (see Joy, Boyle, and Tan [1996]). Using Box-Muller on the quasi-numbers would scramble their order.

⁵Technically, this is weak convergence in a space of stochastic processes. This implies convergence in distribution for the random variables at fixed time points.

⁶These formulas should be understood as using "local references"; x refers to the value at the node we are currently at. Writing things out with "global references" is a notational nightmare, especially if time and step size dependence is made explicit too.

⁷Initially, one would think that the drift function μ could also ruin recombination, but notice how this has cleverly been avoided by letting only the probabilities depend on μ . Advanced readers will see the connection to Girsanov's theorem that says that for Ito processes "a changing probability measure is changing drift, not volatility."

⁸We use the forward equation and not the simpler looking backward equation for the density because, solving the forward equation numerically, we get an approximation in the variable in which the unknown function is a density. This means that we can use known analytic properties (such as the function integrating to 1) to improve the numerical results.

⁹Poulsen [1999, Figure 3] shows how bad a choice is a rectangular box whose integral is 1, although this is also a natural Lebesgue density approximation of a distribution putting all mass at x .

¹⁰Most sources would not include the $1/2$ -factor in the definition, but would have a $(-1)^j$ on the right-hand side. Our definition results in maximal notational compactness of the expressions arising from the analysis.

¹¹For instance:

$\mathcal{A}^0 g = g$, $\mathcal{A}^1 g = \mu_Y g' + \frac{1}{2} g''$ and $\mathcal{A}^2 g = \mu_Y \mu_Y' g' + \mu_Y^2 g'' + \mu_Y g^{(3)} + \frac{1}{2} \mu_Y'' g' + \mu_Y' g'' + \frac{1}{4} g^{(4)}$. It is this step that causes the length of the formulas to increase.

¹²Strictly speaking, we are trying to approximate a function of three variables, $t \sim$ time, $x \sim$ "current state," and $y \sim$ "future state." We give error measurements related to only one of these, namely, y , while keeping the others fixed, since the "error in y " is often the most relevant one. Further our stud-

ies indicate that results are insensitive to the chosen values of t and x .

¹³Further exhibits, as well as finer points about hardware, software, and implementation, are available from the authors upon request.

¹⁴The programs are written in C using double precision. It would be possible to use a high-level language more suited for numerical work, but that would dramatically increase computation times.

REFERENCES

Aït-Sahalia, Y. "Maximum Likelihood Estimation of Discretely Sampled Diffusion Processes: A Closed-Form Approximation Approach." *Econometrica*, Vol. 70 (2002), pp. 223-262.

———. "Transition Densities for Interest Rate and Other Non-linear Diffusions." *Journal of Finance*, Vol. 54, No. 4 (1999), pp. 1361-1395.

Berman, L. "Accelerating Monte Carlo: Quasirandom Sequences and Variance Reduction." *Journal of Computational Finance*, Vol. 1, No. 2 (1998), pp. 79-95.

Bibby, B.M., and M. Sørensen. "Martingale Estimating Functions for Discretely Observed Diffusion Processes." *Bernoulli*, Vol. 1 (1995), pp. 17-39.

Black, F., and M. Scholes. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy*, Vol. 81, No. 3 (1973), pp. 637-654.

Chan, K.C., G.A. Karolyi, F.A. Longstaff, and A.B. Sanders. "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate." *Journal of Finance*, Vol. 47 (1992), pp. 1209-1227.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. "A Theory of the Term Structure of Interest Rates." *Econometrica*, Vol. 53 (1985), pp. 385-407.

Cramér, H. "On Some Classes of Series Used in Mathematical Statistics." In *Proceedings of the Sixth Scandinavian Mathematical Congress*, 1925, pp. 399-425.

Duffie, D. *Dynamic Asset Pricing Theory*, 2nd ed. Princeton: Princeton University Press, 1996.

Duffie, D. and P. Glynn. "Efficient Monte Carlo Simulation of Security Prices." *The Annals of Applied Probability*, Vol. 5, No. 4 (1995), pp. 897-905.

Galanti, S., and A. Jung. "Low-Discrepancy Sequences: Monte Carlo Simulation of Option Prices." *The Journal of Derivatives*, Vol 5, No. 3 (1997), pp. 63-83.

Joy, C., P. Boyle, and K. Tan. "Quasi-Monte Carlo Methods in Numerical Finance." *Management Science*, Vol. 42, No. 6 (1996), pp. 926-938.

Karatzas, I., and S. Shreve. *Brownian Motion and Stochastic Calculus*, 2nd ed. New York: Springer-Verlag, 1992.

Kloeden, P.E., E. Platen, and H. Schurz. *Numerical Solution of SDE Through Computer Experiments*. Heidelberg: Springer, 1994.

Milshtein, G.N. "A Method of Second-Order Accuracy Integration of Stochastic Differential Equations." *Theory of Probability and Its Applications*, 23 (1978), pp. 396-401.

Nelson, D., and K. Ramaswamy. "Simple Binomial Processes as Diffusion Approximations in Financial Models." *Review of Financial Studies*, Vol. 3 (1989), pp. 393-430.

Østerby, O. "The Error of the Crank-Nicolson Method for Linear Parabolic Equations with a Derivative Boundary Condition." Working paper, Department of Computer Science, University of Aarhus, 1998.

Poulsen, R. "Approximate Maximum Likelihood Estimation of Discretely Observed Diffusion Processes." Working Paper No. 29, Centre for Analytical Finance, University of Aarhus, 1999.

Press, W., S. Teukolsky, W.T. Vetterling, and B. Flannery. *Numerical Recipes in C*, 2nd ed. Cambridge: Cambridge University Press, 1992.

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

Strikwerda, J. *Finite Difference Schemes and Partial Differential Equations*. New York: Wadsworth & Brooks/Cole, 1989.

Vasicek, O. "An Equilibrium Characterization of the Term Structure." *Journal of Financial Economics*, Vol. 5 (1977), pp. 177-188.

To order reprints of this article please contact Ajani Malik at amalik@iijournals.com or 212-224-3205.