IMA Journal of Applied Mathematics (2011) **76**, 449−474
doi:10.1093/imamat/hxr016
Advance Access publication on April 16, 2011

Stochastic ordinary differential equations in applied and computational mathematics

DESMOND J. HIGHAM
Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, UK
d.j.higham@strath.ac.uk

[Received on 26 April 2010; revised on 18 January 2011; accepted on 21 February 2011]

Using concrete examples, we discuss the current and potential use of stochastic ordinary differential equations (SDEs) from the perspective of applied and computational mathematics. Assuming only a minimal background knowledge in probability and stochastic processes, we focus on aspects that distinguish SDEs from their deterministic counterparts. To illustrate a multiscale modelling framework, we explain how SDEs arise naturally as diffusion limits in the type of discrete-valued stochastic models used in chemical kinetics, population dynamics and, most topically, systems biology. We outline some key issues in existence, uniqueness and stability that arise when SDEs are used as physical models and point out possible pitfalls. We also discuss the use of numerical methods to simulate trajectories of an SDE and explain how both weak and strong convergence properties are relevant for highly efficient multilevel Monte Carlo simulations. We flag up what we believe to be key topics for future research, focussing especially on non-linear models, parameter estimation, uncertainty quantification, model comparison and multiscale simulation.

Keywords: stochastic computation; modelling; systems biology; finance.

1. Introduction

In the context of modelling physical systems, uncertainty may arise in several ways.

- Directly observable quantities may be subject to measurement error, for example, initial levels in a population model may not be known exactly.
- Parameters that cannot be directly measured may be inferred by calibrating against observations of the system, for example, unknown rate constants in a chemical kinetics model may be fitted against a time series of concentration levels.
- Effects that would be unnecessarily expensive or complicated to measure or model may be summarized stochastically, for example, rather than treating the roll of a die as a non-linear dynamical system, it may be adequate to represent the outcome in terms of a discrete random variable with six possible values.

There are, of course, many ways to introduce randomness into a mathematical model. We focus here on the particular context of ordinary, initial value, stochastic differential equations (SDEs) in Itô form. This class of models is proving popular across a wide range of application areas. In particular, their usefulness in mathematical finance and systems biology has dramatically raised the profile of SDEs. Our aim here is to provide background information and give an overview of some of the key modelling and simulation issues that are likely to have the highest profile over the next few years, with the caveat that we make no attempt to give an exhaustive coverage.

In keeping with the scope and readership of this journal, we have taken an applied mathematics viewpoint. We assume that the reader is familiar with deterministic ordinary differential equations (ODEs) and their numerical approximation but only require a minimal level of familiarity with probability theory (including basic concepts such as normal/Gaussian random variables, probability density functions, independence, expected value, variance and Monte Carlo simulation). We generally focus on a pathwise, or trajectory-based interpretation of an SDE solution, and, where possible, we contrast ideas and results for SDEs with their ODE counterparts. Throughout, the capitalized mathematical font is reserved for random variables, or more generally, stochastic processes.

For further background reading on SDEs we suggest, in roughly increasing order of technical difficulty (Mikosch, 1998; Cyganowski *et al.*, 2002; Mao, 2007; Milstein & Tretyakov, 2004; Kloeden & Platen, 1999).

2. SDEs and their numerical simulation

Given $x_0 \in \mathbb{R}^m$ and a function $f\colon \mathbb{R}^m \to \mathbb{R}^m$, the recurrence relation

$$x_{n+1} = x_n + hf(x_n) \tag{1}$$

is familiar as an Euler approximation to the ODE system $x'(t) = f(x(t))$. Here, the fixed parameter $h > 0$ is called the stepsize, and x_n approximates $x(t_n)$, where $t_n = nh$. Of course, (1) is also an extremely useful analytical tool; by considering the limit $h \to 0$, it is possible to establish existence and uniqueness results for the underlying ODE. In a similar manner, we may interpret an SDE as the limiting process that arises from a discrete-time approximation. To do this, we will give each iterate in (1) an appropriately scaled Gaussian 'kick' producing the Euler–Maruyama iteration

$$X_{n+1} = X_n + hf(X_n) + \sqrt{h}g(X_n)V_n, \tag{2}$$

where

- $g\colon \mathbb{R}^m \to \mathbb{R}^{m\times d}$ is a given function, and
- the $\{V_n\}_{n\geqslant 0}$ are independent vector-valued random variables such that each of the d independent components of V_n has the standard normal distribution.

We see that the magnitude of the random kick in (2) depends upon the current approximation X_n via the value of $g(X_n)$. We also see that the kick scales like $\sqrt{h}$—this turns out to be the right amount of noise to produce limiting trajectories that are continuous but not deterministic.

So, given appropriate functions f and g, and an initial condition $X(0)$, we can think of an SDE solution $X(t)$ as being whatever process arises when we take the $h \to 0$ limit in (2). More precisely, just as in the deterministic case, we can fix t and consider the limit as $h \to 0$ of X_N where $Nh = t$. Of course, for each fixed t, this construction for $X(t)$ leads to a vector-valued random variable, and hence as t varies $X(t)$ is a vector-valued *stochastic process*. In summary, there are three main ingredients for an SDE.

- The function $f\colon \mathbb{R}^m \to \mathbb{R}^m$, called the *drift coefficient*, plays a similar role to the right-hand side of an ODE.

- The function $g\colon \mathbb{R}^m \to \mathbb{R}^{m\times d}$, called the *diffusion coefficient*, governs how the current state of the system affects the size of the noise contribution.
- The initial condition, $X(0)$, may be deterministic, but more generally it is allowed to be a random variable.

The standard notation for specifying such an SDE is

$$\mathrm{d}X(t) = f(X(t))\mathrm{d}t + g(X(t))\mathrm{d}W(t), \quad X(0) \text{ given}, \tag{3}$$

where $W(t)$ is a vector-valued process whose d components represent independent Brownian motions. We will use this notation here, while emphasizing that $\mathrm{d}X(t)$, $\mathrm{d}t$ and $\mathrm{d}W(t)$ have no meaning on their own; we simply regard (3) as a shorthand way of saying that the process $X(t)$ arises from the $h \to 0$ limit in (2).

A simple and very widely used example is given by the scalar ($m = d = 1$) linear case

$$f(x) = ax, \quad g(x) = bx, \tag{4}$$

where a and $b > 0$ are constants. In Fig. 1 we fix $x(0) = 1$, $a = 0.06$, $b = 0.4$ and take $h = 0.01$ in (2). The upper picture in Fig. 1 shows 50 different paths. So, in each case, a Gaussian increment V_n was produced from a call to a standard normal pseudo-random number generator. In this manner, at the final

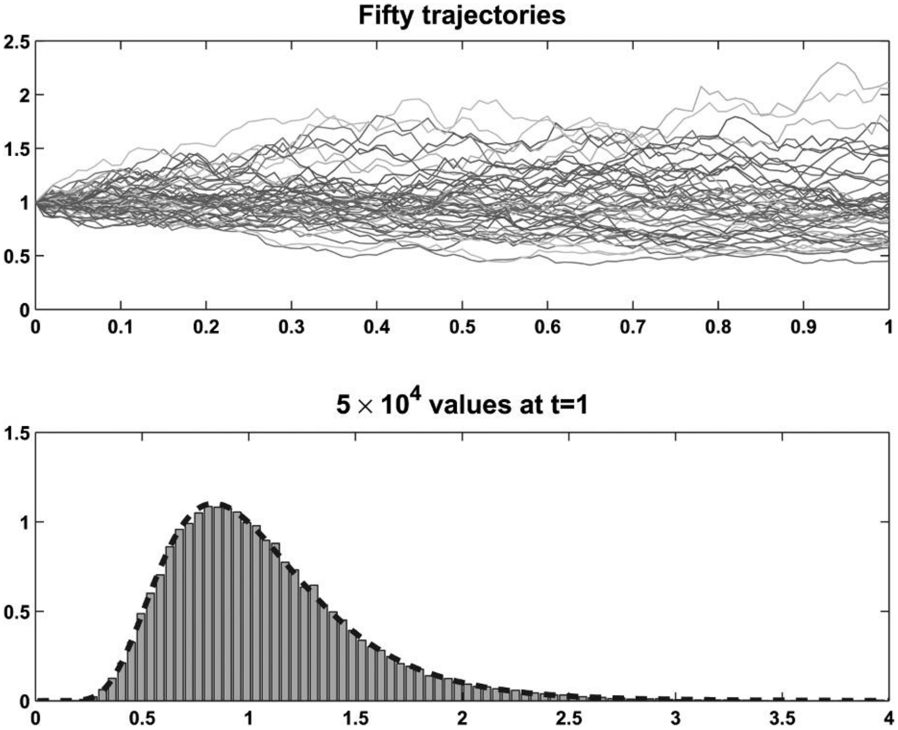

FIG. 1. Upper: fifty paths from the iteration (2) with $x(0) = 1$, $f(x) = 0.06x$, $g(x) = 0.4x$ and $h = 0.01$. Lower: histogram of 5×10^4 binned values for time $t = 1$, with density function (5) superimposed.

time, $t = 1$, each path produces a single number that, in the $h \to 0$ limit, may be regarded as a sample from the distribution of the random variable $X(1)$ describing the SDE solution at $t = 1$. In the lower picture of Fig. 1, we have shaded a histogram for 5×10^4 such samples.

The upper picture shows a trajectory-wise view of an SDE—individual paths are seen to evolve over time. The lower picture applies at a fixed point in time and considers the distribution of values. From the latter perspective, for this simple SDE it can be shown, given a deterministic initial condition, $x(0)$, that the random variable $X(t)$ has a *lognormal* probability density function given by

$$p(y) = \frac{\exp\left(\frac{-\left(\log(y/x(0)) - \left(a - \frac{1}{2}b^2\right)t\right)^2}{2b^2 t}\right)}{yb\sqrt{2\pi t}}, \qquad \text{for } y > 0, \tag{5}$$

and $p(y) = 0$ for $y \leqslant 0$. This density function for $t = 1$ is superimposed in the lower picture of Fig. 1, and we see that it matches the histogram closely.

Of course, the hand-waving arguments leading from (2) to (3) are not valid for arbitrary choices of drift and diffusion coefficient. Generally, the question of existence and uniqueness of solutions for SDEs is more delicate than the ODE case. Most standard texts impose the condition that f and g in (3) are globally Lipschitz—there is assumed to be a constant L such that

$$\|f(u) - f(v)\| \leqslant L\|u - v\|, \tag{6}$$

$$\|g(u) - g(v)\| \leqslant L\|u - v\|, \tag{7}$$

for all $u, v \in \mathbb{R}^m$ (see, for example, Kloeden & Platen, 1999; Mao, 2007). In the ODE case, the right-hand side of a typical ODE model will not satisfy the condition (6), but it is often natural to argue that a *local Lipschitz condition* will hold—a suitable constant $L = L(R)$ will exist for any ball of radius R about the origin. In this way, physical arguments may suggest that the ODE solution will stay bounded, in which case f may be redefined to be zero outside an appropriately large ball, and the local Lipschitz condition can be extended to a global one. This type of reasoning is much harder to justify in the SDE setting. Introducing noise opens up the possibility that trajectories may take arbitrarily large excursions, and establishing existence and uniqueness results is a delicate business, typically hinging on the fact that increasingly large solution values are increasingly less probable.

Similar comments apply when, as in the next section, we analyse numerical methods for simulating SDEs—the textbook global Lipschitz conditions place severe constraints on the class of problems that can be analysed.

3. Stability and convergence of numerical simulations

Numerical methods are traditionally studied in asymptotic regimes. Convergence looks at the error over a finite time interval $[0, T]$ as $h \to 0$ and stability looks at the approximate solution with a fixed h as $t \to \infty$. In both cases, because a random variable is an infinite-dimensional object, the choice of norm is crucial.

The two most widely used convergence concepts are referred to as *weak* and *strong*. Weak error measures how well a numerical method reproduces $\mathbb{E}[X(t)]$ (or, more generally, $\mathbb{E}[\phi(X(t))]$, where $\phi(\cdot)$ is some polynomially bounded function). Under appropriate conditions, which usually include global Lipschitz bounds on the drift and diffusion, the Euler–Maruyama method can be shown to have weak order one so that

$$\sup_{0\leqslant nh\leqslant T} (\mathbb{E}[X(nh)] - \mathbb{E}[X_n]) = \mathrm{O}(h). \tag{8}$$

Strong error, on the other hand, measures the mean of the absolute difference between the two random variables, and Euler–Maruyama achieves only an order of one half in this sense:

$$\mathbb{E}\left[\sup_{0\leqslant nh\leqslant T} |X(nh) - X_n|\right] = \mathrm{O}\left(h^{\frac{1}{2}}\right). \tag{9}$$

More generally, for any $m > 1$ and sufficiently small h there is a constant $C = C(m)$ such that

$$\mathbb{E}\left[\sup_{0\leqslant nh\leqslant T} |X(nh) - X_n|^m\right] \leqslant Ch^{m/2}. \tag{10}$$

Using the Borel–Cantelli lemma, it is possible to pass from strong error to pathwise error. For example, in Kloeden & Neuenkirch (2007), it is shown that given any $\epsilon > 0$, there exists a path-dependent constant $K = K(\epsilon)$ such that, for all sufficiently small h,

$$\sup_{0\leqslant nh\leqslant T} |X(nh) - X_n| \leqslant K(\epsilon)h^{\frac{1}{2}-\epsilon}.$$

In the ODE setting, rates such as $\mathrm{O}(h)$ and $\mathrm{O}\big(h^{\frac{1}{2}}\big)$ might be dismissed as impractical, but for SDE computations, they are frequently tolerated because

- as discussed in Section 6, statistical error generally dominates over discretization error, and
- higher order methods for general SDEs, especially in the strong sense, carry heavy overheads (Kloeden & Platen, 1999).

Hence, although special-purpose higher order methods can be developed for particular circumstances (Anderson & Mattingly, 2011), Euler–Maruyama, or one of its implicit variants, is at the heart of most practical SDE computations.

The linear SDE (4) has proved to be a good starting point for the study of basic long-term behaviour, not least because it gives a natural extension of the classic test problem for numerical ODEs (Hairer & Wanner, 1996). For the SDE itself, there are simple characterizations for mean square stability

$$\lim_{t\to\infty} \mathbb{E}[X(t)^2] = 0 \quad \Leftrightarrow \quad a + \frac{1}{2}b^2 < 0$$

and asymptotic stability

$$\lim_{t\to\infty} |X(t)| = 0, \text{ with probability one} \quad \Leftrightarrow \quad a - \frac{1}{2}b^2 < 0.$$

A typical one-step numerical method produces recurrences of the form

$$X_{n+1} = X_n(p + qV_n), \tag{11}$$

where the coefficients p and q depend on h and on the SDE parameters, a and b. Mean square stability of this discrete iteration is neatly characterized as

$$\lim_{n\to\infty} \mathbb{E}X_n^2 = 0 \quad \Leftrightarrow \quad p^2 + q^2 < 1, \tag{12}$$

but, perhaps surprisingly, the corresponding property of asymptotic stability has a less tractable form; from the strong law of large numbers and the law of the iterated logarithm, we find (Higham, 2000)

$$\lim_{n\to\infty} |X_n| = 0, \text{ with probability one} \quad \Leftrightarrow \quad \mathbb{E}[\log|p + qV_n|] < 0. \tag{13}$$

In Fig. 2, the white bounded area of the p, q plane is the region of asymptotic stability, that is, where the right-hand inequality in (13) holds. The unit circle, marked with a dashed line, is the boundary for mean square stability (12). Both regions are symmetric about the p and q axes, so we only show $p, q > 0$. To emphasize that the two stability concepts are different, we have marked with a cross in Fig. 2 the point $p = q = 1.1$. Here, the iteration is asymptotically stable but not mean square stable—every path must tend to zero as time increases, but for any large time, there are enough 'bad' paths to make the variance huge. Figure 3 shows how one path of $|X_n|$ evolves in this case, with the vertical axis on a logarithmic scale. The iterates decay, albeit far from monotonically.

Given a test problem, we would like our method to reproduce stability for the biggest possible range of stepsizes. A stochastic extension of the trapezoidal rule

$$X_{n+1} = X_n + \frac{1}{2}hf(X_n) + \frac{1}{2}hf(X_{n+1}) + \sqrt{h}\,g(X_n)\,V_n, \tag{14}$$

is easily shown, via (12), to have perfect mean square stability behaviour—given any a and b, and any stepsize h, the method matches the stability/instability of the SDE. For asymptotic stability, however,

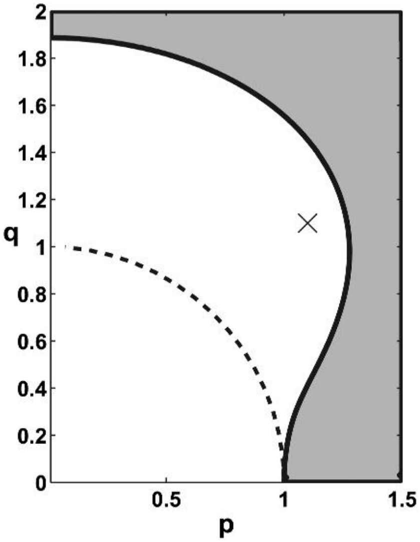

FIG. 2. Stability regions for the iteration (11) for $p, q > 0$. The dashed line along the unit circle is the boundary for mean square stability (12). The solid line is the boundary for asymptotic stability (13). The choice $p = q = 1.1$ used for Fig. 3 is marked with a cross.

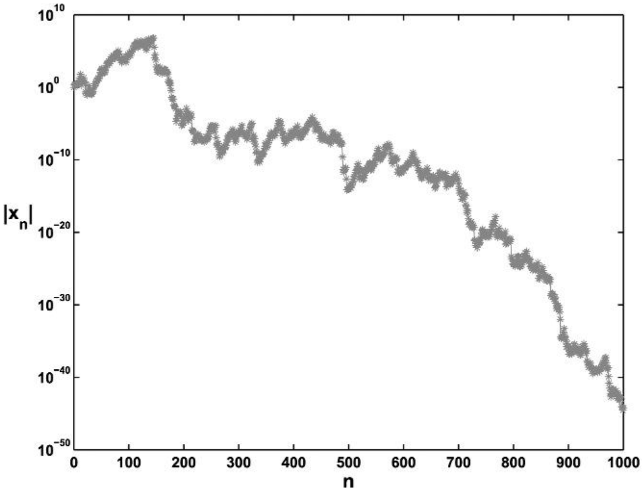

FIG. 3. One instance of the sequence $|X_n|$ for the iteration (11) with $p = q = 1.1$. This process is asymptotically stable but not mean square stable.

analysis via (13) is more awkward, and we are not aware of any general purpose method that can be guaranteed, for all $h > 0$, to preserve asymptotic stability of the SDE.

Although convergence theory under global Lipschitz assumptions and stability theory for a linear test problem give a useful baseline, the study of non-linear SDEs raises new issues and casts doubt on the usefulness of Euler–Maruyama. Several authors Higham *et al.* (2002; 2007), Hutzenthaler *et al.* (2011), Mattingly *et al.* (2002), Milstein & Tretyakov (2005) and Talay (1999) have shown that Euler–Maruyama can fundamentally break down for non-linear and/or long-time computations.

For example, the scalar SDE

$$\mathrm{d}X(t) = -X(t)^3\,\mathrm{d}t + \mathrm{d}W(t),$$

with any deterministic initial condition $X(0)$, has a well-defined solution. However, Hutzenthaler *et al.* (2011, Theorem 1) shows that over any compact interval $[0, T]$, the strong error in an Euler–Maruyama approximation to this SDE blows up as $h \to 0$. Similarly, Higham *et al.* (2007) shows that on the example

$$\mathrm{d}X(t) = (X(t) - X(t)^3)\mathrm{d}t + 2X(t)\mathrm{d}W(t),$$

for which

$$\limsup_{t\to\infty} \frac{1}{t} \log|X(t)| \leqslant -1, \quad \text{with probability } 1,$$

given any $h > 0$ there is a non-zero probability that a path generated by an Euler–Maruyama simulation will blow up as $t \to \infty$. Although these results deal with different types of behaviour, in both cases, their proof relies on the fact that the Gaussian increments used by the numerical method may occasionally perturb the iterates into a region where the non-linear drift has a repulsive effect, and it is clear that any other explicit numerical method can suffer the same fate.

This brings us to a key point. Unlike in the deterministic ODE case, for non-linear SDEs, we introduce implicitness not in the hope of improving efficiency by allowing larger stepsize, but in the hope of obtaining a method that satisfies the fundamental requirements of accuracy and stability.

Although some general results are available for specific non-linear structures, for example, one-sided Lipschitz constants (Higham *et al.*, 2002; Mattingly *et al.*, 2002), many SDE models do not fit into standard categories. Challenges may arise not only through faster than linear growth of the coefficients at infinity but also through unbounded derivatives at the origin—in particular, we will see in Section 5 that square roots arise naturally in models of chemical kinetics. In a specific example motivated by an empirically fitted interest rate model of Ait-Sahalia (1999), strong convergence of specially constructed implicit methods is considered in Szpruch *et al.* (2011) for the problem class

$$\mathrm{d}X(t) = (\alpha_{-1}X(t)^{-1} - \alpha_0 + \alpha_1 X(t) - \alpha_2 X(t)^r)\mathrm{d}t + \sigma X(t)^\rho\, \mathrm{d}W(t),$$

where the α_i are positive constants and $r, \rho > 1$.

With regard to long-time behaviour, the simple fixed point $X(t) \equiv 0$ for the linear test equation (4) is a very special case of an *invariant measure*, and this more general concept can be studied for various classes of non-linear SDE (Mattingly *et al.*, 2002, 2010; Talay, 1999).

4. SDEs as chemical Langevin equations: part 1, motivation

To motivate the use of stochastic models in systems biology, we begin with a simple deterministic example. In Erban *et al.* (2006), a stylized model is given for the levels of two types of protein that are mutually repressive—an increase in the level of protein P_1 inhibits the production of protein P_2 and *vice versa*. Letting $z_1(t)$ and $z_2(t)$ denote the levels of P_1 and P_2 at time t, respectively, a mass action ODE system for this two-gene network takes the form

$$\frac{\mathrm{d}z_1}{\mathrm{d}t} = \frac{1}{1+\kappa z_1}\left(\frac{\gamma}{1+\omega z_2^2} - \delta z_1\right), \tag{15}$$

where the equation for z_2 is found by swapping z_1 and z_2 in (15). Using parameter values $\kappa = 2 \times 10^{-4}$, $\delta = 7.5 \times 10^{-4}$, $\omega = 2 \times 10^{-6}$ and $\gamma = 1.14$, it can be shown that this ODE system has two linearly stable steady states; one has $z_1(t) \equiv z_a \approx 481$ and the other $z_1(t) \equiv z_b \approx 1039$. This type of bistability, predicting that cells may evolve into more than one possible state, is of great biological importance (Hasty *et al.*, 2000). However, for deterministic models such as (15), it may be argued as unrealistic that (a) the cell's fate is completely specified by the initial condition and (b) a cell cannot switch dynamically between states. A major benefit of stochastic models is that they can allow naturally for the scenario where the system spends time in more than one 'attractive' region of state space.

In Fig. 4, which is based on Erban *et al.* (2006, Fig. 5), we show the P_1 protein level arising from a simulation that uses a stochastic analogue of (15). Full details are given later in this section, at this stage, we simply mention that the simulation produces an integer number of P_1 proteins along a discrete set of times t. Every 1000th such value is plotted as a dot in the figure, starting with 600 molecules and running up to time $t = 10^7$. We see that the P_1 level spends time close to each of the two stable steady-state values that exist for the ODE version of the model. It is, of course, possible to study the statistics of the stochastic model further, for example, the typical time for a 'transition' between the two levels may be of interest (Erban *et al.*, 2006).

Bistability, and more general, multistability behaviour for stochastic models is, of course, also of great interest for many other physical and mechanical systems.

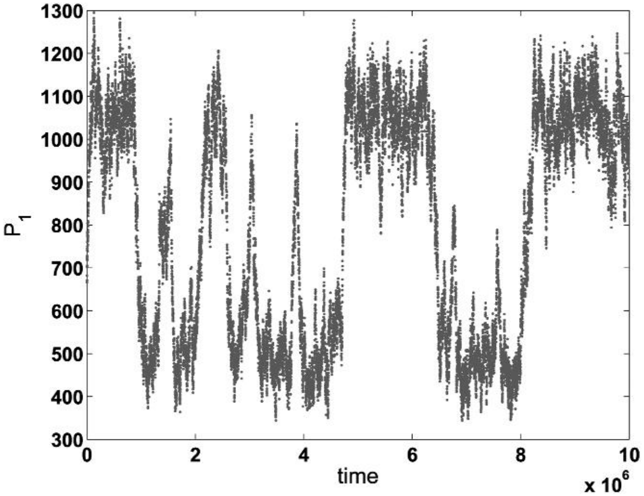

FIG. 4. Number of P_1 molecules over time, from a stochastic model of a genetic toggle switch. The underlying deterministic approximation has stable steady levels at 481 and 1039.

The stochastic simulation in Fig. 4 is based on what is often called the *chemical master equation* (CME) regime, whereas the ODE model (15) corresponds to the mass action or *reaction rate equation* (RRE) setting. Between these two extremes, there is a diffusion limit or *chemical Langevin equation* (CLE) regime that takes the form of an SDE. The CLE regime has the benefit of retaining the stochastic nature of the underlying CME framework, while making simulation and analysis more tractable.

Before discussing the general setting, we will illustrate the main ideas on the extremely simple case of unimolecular decay

$$S \overset{cS}{\rightarrow} \emptyset. \tag{16}$$

Here, we have a single species, S, in our system, and the only event that can take place at any time is that one molecule of S may decay. The rate constant $c > 0$ quantifies the strength of the decay process. We suppose that initially, at time $t = 0$, the number of molecules of S is known to be N. We also note that this system would be called a pure death process in the context of stochastic population modelling (Renshaw, 1991).

In the CME regime, the state of the system at time t is described by a non-negative integer $Z(t)$, representing the number of molecules of S present. Hence, $Z(t)$ may take any of the values $N, N - 1, N - 2, \ldots, 1, 0$. Given that there are $Z(t)$ molecules present at time t, first principle modelling arguments show that the time we must wait before the next reaction takes place (that is, the next time we lose a molecule of S) has an exponential distribution with expected value $1/(cZ(t))$. This is intuitively reasonable—as the number of molecules, $Z(t)$, decreases, we must typically wait longer for the next one to disappear. Similarly, for a system with a smaller rate constant, c, we would typically wait longer between events. Furthermore, the exponential distribution makes the waiting time between events *memoryless*; the chance of the next event occurring within the next second does not depend upon how long ago the last event took place. Because exponentially distributed samples can be easily constructed by log-transforming uniformly distributed samples, it is a very simple matter to compute a path for $Z(t)$.

The following pseudocode summarizes an appropriate algorithm, assuming that the initial state, $Z(0)$, is given.

```
A. Draw a uniform (0,1) pseudo-random sample, $\xi$.
B. Set $\tau = \ln(1/\xi)/(cZ(t))$ to be the waiting time before the next reaction.
C. Update the system to $Z(t+\tau) = Z(t) - 1$ and update the current time $t$ to $t + \tau$.
D. Return to Step A if $Z(t) > 0$ and you wish to continue.
```

In the CLE setting for the reaction (16), we use an SDE to represent the level of species S present at time t. So, at each time t, we have a continuous-valued random variable, $X(t)$. The CLE takes the form of the Itô SDE

$$\mathrm{d}X(t) = -cX(t)\mathrm{d}t - \sqrt{cX(t)}\mathrm{d}W(t). \tag{17}$$

The RRE, or mass action, formulation for (16) is simply the scalar ODE $\mathrm{d}z(t)/\mathrm{d}t = -cz(t)$, where $z(t)$ is a deterministic real-valued quantity representing the amount of S present at time t.

In Fig. 5, we illustrate the three regimes in the case of ten initial molecules. (The CLE was simulated numerically using Euler–Maruyama.) It is immediately apparent that the CLE path does not respect the inherent monotonicity of this simple reaction. Unlike the RRE solution, however, any CLE path will, eventually, attain the value zero. Figures 6 and 7 repeat the exercise with 50 and 200 initial molecules, respectively. We see that the fluctuations are less significant when the molecule count is high—this idea will be formalized shortly when we consider the thermodynamic limit.

In the CME regime for (16), at every time t the state $Z(t)$ is a random variable with a discrete set of possible values $0, 1, 2, \ldots, N$. We may then let $p_i(t)$ denote the probability that $Z(t) = i$. It follows that $\{p_i(t)\}_{i=0}^{N}$ satisfy an ODE, or master equation, of the form

$$\frac{\mathrm{d}}{\mathrm{d}t} p_i(t) = c(i+1)p_{i+1}(t) - cip_i(t), \quad \text{for } i = N-1, N-2, \ldots, 0, \tag{18}$$

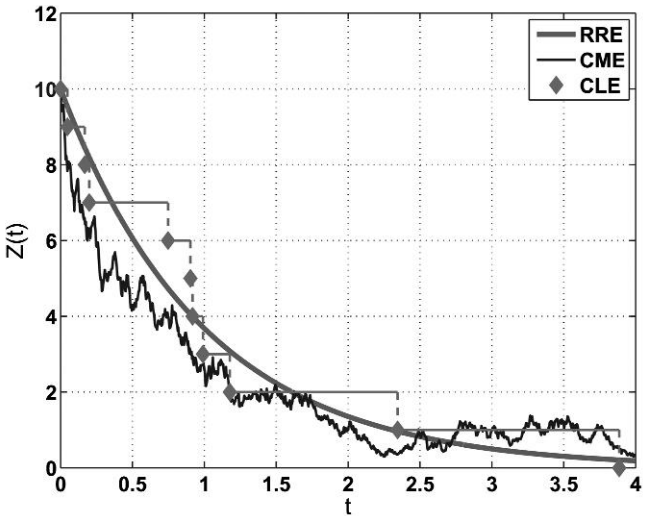

FIG. 5. Simulations of the simple reaction (16), starting with 10 molecules.

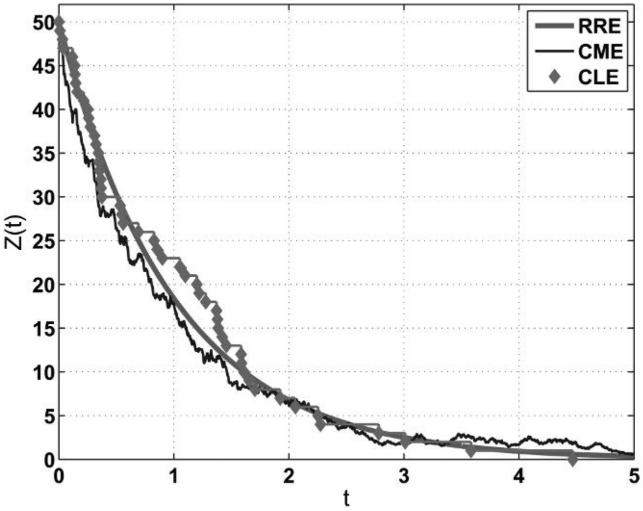

FIG. 6. Simulations of the simple reaction (16), starting with 50 molecules.

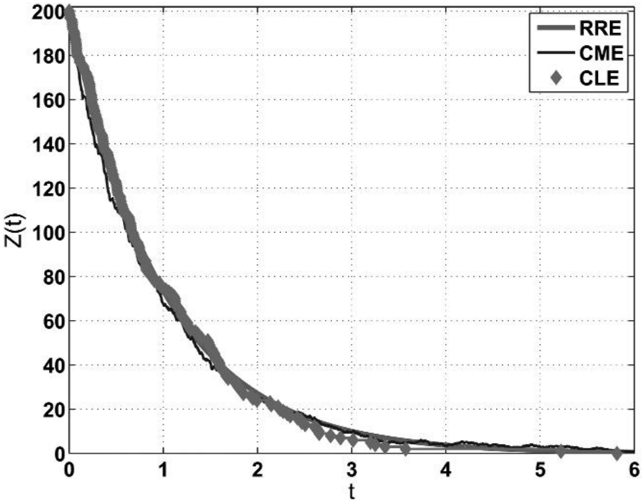

FIG. 7. Simulations of the simple reaction (16), starting with 200 molecules.

where $P_{N+1}(t)$ is taken to be zero. This has the intuitive interpretation that the rate of change of $p_i(t)$ has

- a positive contribution $c(i+1)p_{i+1}(t)$, which corresponds to the fact that we enter state i via one decay from state $i+1$, and
- a negative contribution $-cip_i(t)$ due to the fact that we leave state i when a decay takes place.

The linear ODE system (18) has solution

$$p_i(t) = \frac{N!}{i!(N-i)!}\mathrm{e}^{-cit}(1-\mathrm{e}^{-ct})^{N-i}, \quad \text{for } i = 0, 1, 2, \ldots, N, \tag{19}$$

and it follows that the mean, $\mathbb{E}[Z(t)]$ and variance $\mathsf{var}[Z(t)]$ have the form

$$\mathbb{E}[Z(t)] = N\,\mathrm{e}^{-ct} \quad \text{and} \quad \mathsf{var}[Z(t)] = N\,\mathrm{e}^{-ct}(1-\mathrm{e}^{-ct}). \tag{20}$$

For the CLE (17), because the drift coefficient $-cX(t)$ is linear, it follows immediately that $\mathbb{E}[X(t)]$ satisfies the ODE that arises when the noise is switched off, giving

$$\mathbb{E}[X(t)] = N\,\mathrm{e}^{-ct}. \tag{21}$$

To find the second moment, we may apply Itô's lemma, see, for example, Mao (2007), to get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X(t)^2] = -2c\mathbb{E}[X(t)^2] + c\mathbb{E}[X(t)],$$

from which it follows that

$$\mathsf{var}[X(t)] = N\,\mathrm{e}^{-ct}(1-\mathrm{e}^{-ct}). \tag{22}$$

So the CLE reproduces the mean and variance of the CME.

The RRE matches the mean of the CME, that is, $z(t) = \mathbb{E}[Z(t)] = N\,\mathrm{e}^{-ct}$. Being deterministic, $z(t)$ of course has zero variance.

Studying the first and second moments in this way outlines one sense in which the CLE may be regarded as an intermediate model that approximates the CME more accurately than the RRE. In the next section, we look at this issue in more detail.

5. SDEs as CLE: part 2, theory and challenges

Suppose we have a general system with R chemical species, $S_1, S_2, \ldots, S_R$, taking part in M different chemical reactions. In the CME formulation, we then have a state vector $Z(t) \in \mathbb{R}^R$ whose ith component denotes the number of molecules of S_i present at time t. In this setting, unlike in the molecular dynamics regime (Leimkuhler & Reich, 2005), we are not concerned with the location or velocity of each molecule, we simply wish to record the total number for each species. Having settled on this level of detail, we must accept that the most accurate description of how the system evolves must be stochastic. After making some reasonable assumptions (such as a fixed volume for the system and a constant temperature) Gillespie (1976, 1977) used first principle modelling arguments to derive the CME for $Z(t)$. For each $1 \leqslant j \leqslant M$, the CME involves

- a *stoichiometric vector*, $\nu_j \in \mathbb{R}^R$, and
- a *propensity function*, $a_j(Z(t))$,

such that the jth reaction takes place over the infinitesimal interval $[t, t+\mathrm{d}t)$ with probability $a_j(Z(t))\mathrm{d}t$ and causes the change $Z(t) \mapsto Z(t) + \nu_j$ to the state vector. Gillespie showed how to derive appropriate propensity functions for standard chemical reactions.

Letting $P(z,t)$ denote the probability that $Z(t) = z$, the CME is given by the ODE system

$$\frac{\mathrm{d}P(z,t)}{\mathrm{d}t} = \sum_{j=1}^{M} (a_j(z-\nu_j)P(z-\nu_j,t) - a_j(z)P(z,t)). \tag{23}$$

We note that the same form of ODE has been derived in many other modelling contexts, notably population dynamics (Renshaw, 1991), and is often referred to as the *forward Kolmogorov equation*.

For the simple reaction (16), we have $R = 1$ species, $\nu_1 = -1$ and $a_1(x) = cx$, and we see that (23) reduces to (18).

Generally, since z ranges over the set of all possible systems states, the CME represents a massive (albeit linear, constant coefficient) ODE system that is too large to compute with and visualize; although progress is being made for some non-trivial examples (Jahnke, 2010).

As an alternative to computing $P(z,t)$ directly, Gillespie showed that it is possible to compute sample paths that respect these probabilities. In this approach, on each step, we draw two random numbers. One is used to choose a waiting time until the next reaction takes place—this is exponentially distributed with mean given by the inverse of the sum of the values of propensity functions, so the higher the propensities the shorter the typical waiting times. The other is used to choose which of the M reactions to fire. The chance that reaction j fires is proportional to its propensity. Overall, the resulting algorithm can be summarized very simply in the following pseudocode, given an initial state $Z(0)$, which generalizes the special case outlined in Section 4 for unimolecular decay.

```
1. Evaluate $\{a_k(Z(t))\}_{k=1}^{M}$ and $a_{\mathrm{sum}}(Z(t)) := \sum_{k=1}^{M} a_k(Z(t))$.
2. Draw two independent uniform (0,1) random numbers, $\xi_1$ and $\xi_2$.
3. Set $j$ to be the smallest integer satisfying $\sum_{k=1}^{j} a_k(Z(t)) > \xi_1 a_{\mathrm{sum}}(Z(t))$.
4. Set $\tau = \ln(1/\xi_2)/a_{\mathrm{sum}}(Z(t))$.
5. Set $Z(t+\tau) = Z(t) + \nu_j$ and update $t$ to $t + \tau$.
6. Return to Step 1 or terminate.
```

In Fig. 4, we used this algorithm with $R = 2$ species, $M = 4$ reactions, stoichiometric vectors of the form

$$\nu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \nu_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \nu_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \nu_4 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

and propensity functions

$$\begin{aligned} a_1(z) &= \frac{\gamma}{(1+\kappa z_1)(1+\omega z_2^2)}, \\ a_2(z) &= \frac{\delta z_1}{1+\kappa z_1}, \\ a_3(z) &= \frac{\gamma}{(1+\kappa z_2)(1+\omega z_1^2)}, \\ a_4(z) &= \frac{\delta z_2}{1+\kappa z_2}. \end{aligned}$$

We should also mention that although, as any search engine will reveal, Gillespie's algorithm is now extremely well cited in the chemistry and biochemistry fields, other application areas use similar ideas under different names, including the residence-time algorithm (Cox and Miller, 1965), kinetic Monte Carlo (Young & Elcock, 1966) and, more generally, discrete event simulation and Petri nets (Wilkinson, 2006).

Because Gillespie's algorithm faithfully reproduces the statistics of the CME, it is forced to take account of every reaction along a path—the propensity functions must be reevaluated at each new state. If we make an approximation by freezing the propensity functions over some time period, τ, then we can argue that the number of type j reactions taking place arises from a simple counting process and will follow a Poisson distribution with parameter $a_j(Z(t))\tau$. (A Poisson random variable with parameter $\lambda > 0$ takes the value i with probability $\mathrm{e}^{-\lambda}\lambda^i/(i!)$ for $i = 0, 1, 2, \ldots$.) If we further argue that $a_j(Z(t))\tau$ is large, then this Poisson update to the state vector can be approximated by a Gaussian with the same mean and variance. This leads us to the recurrence

$$Y(t+\tau) = Y(t) + \tau \sum_{j=1}^{M} \nu_j a_j(Y(t)) + \sqrt{\tau} \sum_{j=1}^{M} \nu_j \sqrt{a_j(Y(t))}\xi_j, \tag{24}$$

where the ξ_j are independent standard Gaussians and hence each $Y(t)$ is a real-valued random variable. We see from (2) that this has the form of an Euler–Maruyama iteration, and hence, for small τ, we could approximate this system with the SDE

$$\mathrm{d}X(t) = \sum_{j=1}^{M} \nu_j a_j(X(t))\mathrm{d}t + \sum_{j=1}^{M} \nu_j \sqrt{a_j(X(t))}\mathrm{d}W_j(t). \tag{25}$$

This is the CLE model for the chemical system. We saw the simple case (17) for the pure decay reaction (16).

We note also that the iterations of the type (24) are of independent practical interest (see for, example, Anderson *et al.*, 2011; Gillespie, 2001).

To discuss the sense in which the CLE approximates the CME, it is usual to rescale the process $Z(t)$ to $\widehat{Z}(t) = Z(t)/V$, where $V \gg 1$. Typically, V is regarded as the product of the Avagadro constant and the volume in litres, so that $\widehat{Z}(t)$ measures moles per litre. If we similarly scale the CLE solution to $\widehat{X}(t) = X(t)/V$, then, under the assumption that the propensity functions satisfy $a_j(Vx) = \mathrm{O}(V)$ as $V \to \infty$, which holds for standard chemical kinetics, Kurtz (1981) has shown that over a finite time interval $[0, T]$, the largest deviation of $\widehat{Z}(t) - \widehat{X}(t)$ is typically $\mathrm{O}(\log(V)/V)$ (see, also, for example, Anderson *et al.*, 2011; Ball *et al.*, 2006, for more details).

As $V \to \infty$, which is the so called *thermodynamic limit* that we illustrated in Figs 5–7, the deterministic RRE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_j a_j(x(t)), \tag{26}$$

also approximates the discrete stochastic model in the sense that $\widehat{x}(t) = x(t)/V$ matches $\widehat{Z}(t)$ pathwise to $\mathrm{O}(1/\sqrt{V})$.

We have outlined how the CLE can be derived from the CME under certain modelling assumptions and mentioned accuracy over compact time intervals in the thermodynamic $V \to \infty$ limit. It is perhaps not surprising, however, that issues arise when the modelling assumptions are not valid and when long-time behaviour is studied. For an illustration, we may use the simple reversible reaction example,

$$S_1 \underset{k_2}{\overset{k_1}{\rightleftarrows}} S_2, \tag{27}$$

which has stoichiometric vectors

$$\nu_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \nu_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

and propensity functions $a_1(z) = k_1 z_2$, $a_2(z) = k_2 z_1$. Since the only possible events are that a molecule of S_1 converts to a molecule of S_2 , or *vice versa*, it is clear that in the CME framework the total number of molecules is preserved. Further, if we start with a deterministic number $Z_1(t) + Z_2(t) = K$ of molecules, then both $Z_1(t)$ and $Z_2(t)$ must take integer values in the range $\{0, 1, 2, \ldots, K-2, K-1, K\}$.

The CLE for this model has the form

$$\mathrm{d}X_1(t) = (-k_1 X_1(t) + k_2 X_2(t))\mathrm{d}t - \sqrt{k_1 X_1(t)}\mathrm{d}W_1(t) + \sqrt{k_2 X_2(t)}\mathrm{d}W_2(t), \tag{28}$$

$$\mathrm{d}X_2(t) = (k_1 X_1(t) - k_2 X_2(t))\mathrm{d}t + \sqrt{k_1 X_1(t)}\mathrm{d}W_1(t) - \sqrt{k_2 X_2(t)}\mathrm{d}W_2(t). \tag{29}$$

Looking at (28), we see that for $X_1(t)$ close to zero, the right-hand side has a deterministic contribution $k_2 X_2(t)\mathrm{d}t$ pushing $X_1(t)$ back into the positive orthant, but it also has a stochastic contribution $\sqrt{k_2 X_2(t)}\mathrm{d}W_2(t)$ which is equally likely to have a positive or negative effect. Similar comments apply to the case $X_2(t) \approx 0$, and, overall, solutions to (28) and (29) cannot be guaranteed to remain positive. If the molecule count for species X_1 or X_2 becomes small—in which case, the assumptions used to derive the CLE are invalid—the SDE model breaks down because the diffusion coefficients involve square roots of negative arguments. For this reason, analysis of general CLE systems requires care, and modifications to the basic CLE may be required simply to produce a well-defined mathematical object (Szpruch & Higham, 2010).

Wilkie & Wong (2008) noted that the CLE can produce negative concentrations and suggested a fix that involves deleting the offending diffusion coefficients. For the simple example (27), their modified CLE takes the form

$$\mathrm{d}\overline{X}_1(t) = (-k_1 \overline{X}_1(t) + k_2 \overline{X}_2(t))\mathrm{d}t - \sqrt{k_1 \overline{X}_1(t)}\mathrm{d}W_1(t), \tag{30}$$

$$\mathrm{d}\overline{X}_2(t) \ = (k_1 \overline{X}_1(t) - k_2 \overline{X}_2(t))\mathrm{d}t - \sqrt{k_2 \overline{X}_2(t)}\mathrm{d}W_2(t). \tag{31}$$

However, we would argue that all four diffusion terms in (28) and (29) have a role to play in capturing the fluctuations of the underlying Poisson processes about their mean, and we would not recommend making a global change to fix a difficulty that is localized to the boundary.

Because the propensity functions in this example are linear, we can study the issue further by obtaining closed-form ODEs for the evolution of the moments. In the CME framework, the scaled, discrete-valued, process $\widehat{Z}(t)$ has moments that evolve according to the linear ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \mathbb{E}[\widehat{Z}_1] \\ \mathbb{E}[\widehat{Z}_2] \\ \mathbb{E}[(\widehat{Z}_1)^2] \\ \mathbb{E}[(\widehat{Z}_2)^2] \\ \mathbb{E}[\widehat{Z}_1\widehat{Z}_2] \end{bmatrix} = \begin{bmatrix} -k_1 & k_2 & 0 & 0 & 0 \\ k_1 & -k_2 & 0 & 0 & 0 \\ k_1/V & k_2/V & -2k_1 & 0 & 2k_2 \\ k_1/V & k_2/V & 0 & -2k_2 & 2k_1 \\ -k_1/V & -k_2/V & k_1 & k_2 & -(k_1+k_2) \end{bmatrix} \begin{bmatrix} \mathbb{E}[\widehat{Z}_1] \\ \mathbb{E}[\widehat{Z}_2] \\ \mathbb{E}[(\widehat{Z}_1)^2] \\ \mathbb{E}[(\widehat{Z}_2)^2] \\ \mathbb{E}[\widehat{Z}_1\widehat{Z}_2] \end{bmatrix}, \tag{32}$$

see, for example, Gadgil *et al.* (2005) for details of how to derive these relations. For the modified Langevin process (30) and (31), we may apply Itô's lemma (Mao, 2007) to the functions X_1^2, X_2^2, $X_1 X_2$

and then take expectations, to obtain, in scaled form, $\widehat{\overline{X}}(t) = \overline{X}(t)/V$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \mathbb{E}[\widehat{\overline{X}}_1] \\ \mathbb{E}[\widehat{\overline{X}}_2] \\ \mathbb{E}[(\widehat{\overline{X}}_1)^2] \\ \mathbb{E}[(\widehat{\overline{X}}_2)^2] \\ \mathbb{E}[\widehat{\overline{X}}_1\widehat{\overline{X}}_2] \end{bmatrix} = \begin{bmatrix} -k_1 & k_2 & 0 & 0 & 0 \\ k_1 & -k_2 & 0 & 0 & 0 \\ k_1/V & \underline{0} & -2k_1 & 0 & 2k_2 \\ \underline{0} & k_2/V & 0 & -2k_2 & 2k_1 \\ \underline{0} & \underline{0} & k_1 & k_2 & -(k_1+k_2) \end{bmatrix} \begin{bmatrix} \mathbb{E}[\widehat{\overline{X}}_1] \\ \mathbb{E}[\widehat{\overline{X}}_2] \\ \mathbb{E}[(\widehat{\overline{X}}_1)^2] \\ \mathbb{E}[(\widehat{\overline{X}}_2)^2] \\ \mathbb{E}[\widehat{\overline{X}}_1\widehat{\overline{X}}_2] \end{bmatrix}. \tag{33}$$

In (33), we have underlined the zero coefficients in the ODE Jacobian that replace the non-zeros in the master equation version (32). Making such an $\mathrm{O}(1/V)$ change to the entries will generally cause an $\mathrm{O}(1/V)$ change in the ODE solution, over any finite time interval.

In Fig. 8, we show computations for the case $k_1 = k_2 = 1$, with deterministic initial data $X_1^V(0) = 4$ and $X_2^V(0) = 1$. The picture on the left uses $V = 1$ and the picture on the right uses $V = 10$. We have plotted the evolution of the second moment of the first species. The solid curve shows $\mathbb{E}[(\widehat{Z}_1)^2]$, for the master equation formulation, and the thick dashed line shows $\mathbb{E}[(\widehat{\overline{X}}_1)^2]$ for the modified Langevin. The thinner dashed curve, which is the same in both pictures, shows the corresponding deterministic curve for the mass action ODE. We see that the modified Langevin (30) and (31) is no more accurate than

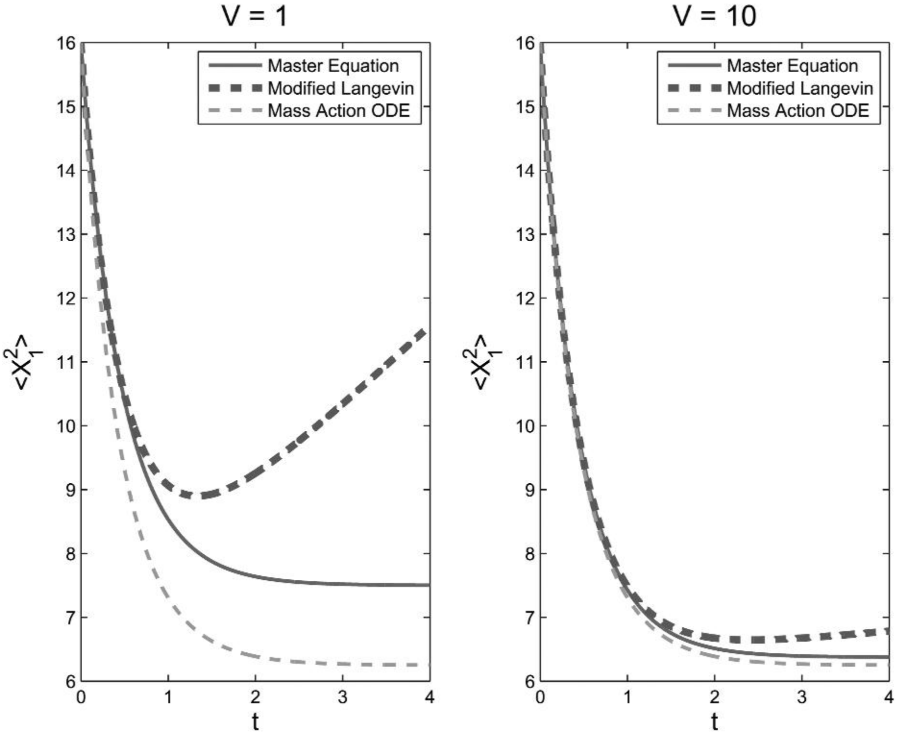

FIG. 8. Second moment of X_1 in the reversible isometry (27) for chemical master equation (solid), modified Langevin (thick dashed) and mass action ODE (thin dashed). Left: $V = 1$. Right: $V = 10$.

the simple mass action ODE in terms of reproducing the second moment from the master equation. We repeated the computations for a range of V values and recorded the error in the second moment at time $t = 4$. For the modified Langevin, we obtained errors that scaled like $\mathbb{E}[(\widehat{Z}_1)^2] - \mathbb{E}[(\widehat{\widehat{X}}_1)^2] = -4.06/V$ and for the mass action ODE this became $1.25/V$.

Using Itô's lemma on the original Langevin equation (28) and (29), we recover the exact moment equation (32)[1]. This makes it clear that the discrepancies underlined in (33) are a direct consequence of setting particular noise terms to zero—a global perturbation to the Langevin equation has reduced its accuracy down to that of the mass action ODE.

In general, dealing systematically with the multiscale interface between discrete-valued stochastic, real-valued stochastic and real-valued deterministic models in order to make large-scale modelling and computation a feasible proposition remains a very active and challenging field that naturally leads into mixed, or hybrid, models that couple or extend the concept of an SDE (Anderson *et al.*, 2011; Ball *et al.*, 2006; Cao *et al.*, 2005; E *et al.*, 2005; Intep *et al.*, 2009; Leier *et al.*, 2008).

6. Monte Carlo Simulations

Most of the computations that are performed on stochastic models can be cast in terms of a Monte Carlo simulation to approximate an expected value Ripley (1987). In the case where SDEs are simulated there is an inherent discretization error—each sample that we compute has a built-in bias because we do not solve the SDE exactly.

For simplicity of exposition, we will suppose in this section that the SDE is scalar—the conclusions hold for general systems. Suppose we wish to find the expected value of some function of the final time solution of this scalar SDE; say $\mathbb{E}[F(X(T))]$, where $F\colon \mathbb{R} \to \mathbb{R}$ is assumed to be globally Lipschitz and $X(T)$ is the final time solution. For example, in the case where we wish to value a European call option in mathematical finance Higham (2004) the SDE models the dynamics of an asset, under a risk-neutral measure, and we have a piecewise linear 'hockey-stick' pay-off function $F(x) = \max(x - E, 0)$, where E is the exercise price.

Given a stepsize h such that $Kh = T$, we could apply the Euler–Maruyama method (2) N times to get approximate samples $\{X_K^{[i]}\}_{i=1}^{N}$ from the distribution of $X(T)$. Here, $X_K^{[i]}$ denotes the final time Euler–Maruyama approximation from the ith path. Our computed approximation to $\mathbb{E}[X(T)]$ would then be the sample mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} X_K^{[i]}.$$

The overall error splits naturally into two terms

$$\begin{aligned} \mathbb{E}[X(T)] - \mu &= \mathbb{E}[X(T) - X_K + X_K] - \mu \\ &= \mathbb{E}[X(T) - X_K] + \mathbb{E}[X_K] - \mu. \end{aligned}$$

The term $\mathbb{E}[X(T) - X_K]$ represents the bias from the discretization error, and the weak error result (8) for Euler–Maruyama shows that this is $\mathrm{O}(h)$. The term $\mathbb{E}[X_K] - \mu$ represents the inherent statistical error associated with Monte Carlo, and, from the central limit theorem, the width of a confidence interval

[1] This exactness is a consequence of the linearity in the propensity functions; generally the error in the moments would be $\mathrm{O}(1/V^2)$.

(to be concrete, we will assume that a 95% confidence interval is required) scales like $\mathrm{O}(1/\sqrt{N})$. Hence, allowing for both sources of error, we have an overall confidence interval of width $\mathrm{O}(h) + \mathrm{O}(1/\sqrt{N})$. Suppose that we wish to obtain a prescribed target accuracy of ε. Then, to avoid unnecessary computation, it makes sense to balance the two terms, so that h scales like ε and N scales like ε^{-2}. If we measure computational cost in terms of either

- the number of pseudo-random numbers generated or
- the number of drift and diffusion coefficient evaluations required,

then the cost is proportional to the product of the number of steps per path, $1/h$, and the number of paths, N. Hence, the cost to obtain a confidence interval width bounded by ε scales like $N/h = \varepsilon^{-3}$.

This conclusion, that for Monte Carlo/SDE simulations the cost varies inversely with the third power of the required accuracy, appears in many standard references.

An obvious way to improve the complexity would be to use a numerical method with a higher weak order. For example, under extra conditions on the SDE coefficients, Talay & Tubaro (1990) showed that an extrapolated version of Euler–Maruyama could be used to increase the weak error rate to $\mathrm{O}(h^2)$. This would improve the computational complexity to $\mathrm{O}(\varepsilon^{-2.5})$.

However, a radically different approach that gives a complexity of $\mathrm{O}(\varepsilon^{-2}(\log \varepsilon)^2)$ was recently put forward by Giles (2008), and it is this extremely promising *multilevel Monte Carlo* (MLMC) technique that we describe here. We can motivate the approach heuristically by noting that is not necessary to compute all paths with the same stepsize h. Because a smaller h is more expensive, it might be beneficial to compute many cheap, low-resolution samples, and then use a few high-resolution paths to fill in the high-frequency detail. More precisely, Giles proposed a hierarchy of discretisation scales in a manner reminiscent of a multigrid computation for a partial differential equation. Before outlining and justifying the main ideas, we wish to emphasize that

- the technique does not rely on a special SDE discretization scheme or a special structure for the SDE—the standard Euler–Maruyama method can be used and the analysis simply exploits its basic weak and strong convergence properties,
- although our aim is to compute an expected value, the technique relies on both the weak *and the strong* convergence behaviour of the numerical method.

In its simplest form, MLMC uses a range of stepsizes of the form $h_l = 2^{-l}T$ for levels $l = 0, 1, 2, \ldots, L$. The number of levels L is chosen so that

$$L = \frac{\log(\varepsilon^{-1})}{\log(2)}. \tag{34}$$

This ensures that at the finest level, L, we have stepsize $h_L = \mathrm{O}(\varepsilon)$. So the bias at this level has the appropriate size.

Now, we let the random variable P_l denote the result of applying Euler–Maruyama with stepsize h_l in order to approximate the pay-off $F(X(T))$. Rather than going for $\mathbb{E}[P_L]$ directly, we will make use of the trivial identity

$$\mathbb{E}[P_L] = \mathbb{E}[P_0] + \sum_{l=1}^{L} \mathbb{E}[P_l - P_{l-1}]$$

and estimate separately the terms on the right-hand side. To do this, at level 0, we will use N_0 paths in order to form the sample average

$$Y_0 = \frac{1}{N_0} \sum_{i=1}^{N_0} P_0^{[i]}, \tag{35}$$

and generally for level $l \geqslant 1$, we will use N_l paths in order to compute

$$Y_l = \frac{1}{N_l} \sum_{i=1}^{N_l} (P_l^{[i]} - P_{l-1}^{[i]}), \tag{36}$$

so that our overall estimator is $Y := Y_0 + \sum_{l=1}^{L} Y_l$. We emphasize here that $P_l^{[i]}$ and $P_{l-1}^{[i]}$ are computed from the same Brownian path. In other words, suppose that we are currently at time t_n, where n is even. If the Euler–Maruyama computation with stepsize h_l uses random increments $\sqrt{h_l}\,\xi_n^{[i]}$ and $\sqrt{h_l}\,\xi_{n+1}^{[i]}$ during the two steps that update to time $t_n + 2h_l$, then the accompanying Euler–Maruyama computation with stepsize $h_{l-1} = 2h_l$ uses $\sqrt{h_l}\xi_n^{[i]} + \sqrt{h_l}\xi_{n+1}^{[i]}$. Figure 9 illustrates this scenario. For each path i, we use independent random increments, and these increments are also independent across different levels—so the pseudo-random numbers are not reused.

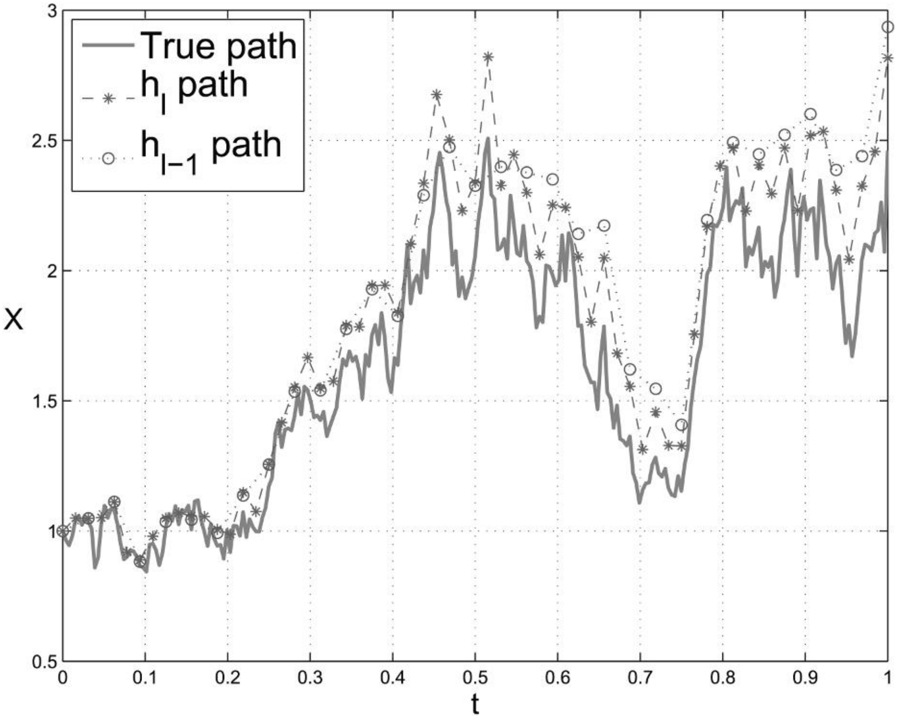

FIG. 9. An illustration of how the estimator (36) is constructed by applying Euler–Maruyama over the same Brownian path with two different stepsizes, h_l and $h_{l-1} = 2h_l$.

It remains to work out how many paths are required at each level in order to reduce the variance in the overall estimate to $\mathsf{var}[Y] = \mathrm{O}(\varepsilon^2)$, so that the final confidence interval has width of $\mathrm{O}(\varepsilon)$, and then to check the resulting computational complexity.

Using the basic inequality $\mathsf{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \leqslant \mathbb{E}[X^2]$ and the global Lipschitz property of F, we have

$$\mathsf{var}[P_l - F(X(T))] \leqslant \mathbb{E}[(P_l - F(X(T)))^2] = \mathrm{O}(\mathbb{E}[(X_K - X(T))^2]).$$

It then follows from the *strong* convergence property (10) of Euler–Maruyama, with $m = 2$, that

$$\mathsf{var}[P_l - F(X(T))] = \mathrm{O}(h_l).$$

Using this inequality along with the appropriate triangle inequality ($\|X + Y\|_2 \leqslant \|X\|_2 + \|Y\|_2$ for $\|X\|_2 := \sqrt{\mathbb{E}[X^2]}$), we find that

$$\mathsf{var}[P_l - P_{l-1}] \leqslant (\sqrt{\mathsf{var}[P_l - F(X(T))]} + \sqrt{\mathsf{var}[P_{l-1} - F(X(T))]})^2 = \mathrm{O}(h_l).$$

It follows that Y_l in (36) has variance of $\mathrm{O}(h_l/N_l)$. Now, since the computations at each level are independent, the overall variance of the estimator Y expands as

$$\mathsf{var}[Y] = \mathsf{var}[Y_0] + \sum_{l=1}^{L} \mathsf{var}[Y_l] = \mathsf{var}[Y_0] + \sum_{l=1}^{L} \mathrm{O}(h_l/N_l).$$

The choice

$$N_l = \mathrm{O}(\varepsilon^{-2} L h_l)$$

is then seen to produce the required overall variance of $\mathsf{var}[Y] = \mathrm{O}(\varepsilon^2)$.

Now the computational complexity of this algorithm is given by the sum over all levels of the product of ‘cost per step’ and ‘number of steps’, which becomes

$$\sum_{l=0}^{L} N_l h_L^{-1} = \sum_{l=0}^{L} \varepsilon^{-2} L h_l h_l^{-1} = L^2 \varepsilon^{-2}.$$

From (34), this leads to a complexity of $\mathrm{O}(\varepsilon^{-2}(\log \varepsilon)^2)$.

In addition to proposing and justifying MLMC, Giles (2008) also implemented a practical version that was seen to deliver the improved complexity on realistic problems in option valuation. Subsequent work on this multilevel approach has looked at

- numerical methods with higher weak and strong order (Giles, 2007),
- various classes of ‘pay-off’ functions F that are not globally Lipschitz and may even depend upon $X(t)$ along the whole path $0 \leqslant t \leqslant T$, for example, the case of barrier options (Avikainen, 2009; Giles *et al.*, 2009),
- MLMC combined with quasi-Monte Carlo methods that improve the statistical component of the complexity (Giles & Waterhouse, 2009).

To put MLMC in context, we emphasize that for standard Monte Carlo simulations

- samples are assumed to be exact, and
- *variance reduction* techniques to speed up the computations typically exploit problem-dependent structures.

By contrast, MLMC applies to the scenario where the samples have a built-in bias arising from an SDE discretization and requires no extra knowledge of the problem structure. In the cases where it has been shown to work, it makes the cost of the SDE simulation negligible—the asymptotic complexity is effectively reduced to the level that would remain if we were able to evaluate the SDE solution exactly. There are, of course, many promising avenues for this remarkable idea not only for SDE simulations but also within the broader context of multiscale modelling and simulation.

7. Model calibration and inference problems

Any mathematical model can only be an approximate description of a physical system. Moreover, it is often the case that some or all the parameters and initial conditions are unknown, and hence must be inferred from experimental measurements. In the SDE case, where the model itself is stochastic, it is natural to quantify this uncertainty by using statistical tools.

We give here a very simple illustration of a Bayesian approach to parameter estimation. We refer to Jaynes (2003) and Sivia & Skilling (2006) for general background information on Bayesian inference, while noting that it is currently something of a novelty in the applied mathematics literature; we recommend the recent survey (Stuart, 2010) for further details about how these topics intersect. We will consider a financial setting where daily observations of an asset are available. Suppose the asset, $S(t)$, is modelled by the simple linear SDE (4)—this assumption is at the heart of the classic Black–Scholes theory for financial option valuation (Higham, 2004). Setting $\Delta t = 1$ day, the asset values $\{S(i\,\Delta t)\}$ may be converted into log-return data

$$R_i = \log\left(\frac{S(i\,\Delta t)}{S((i-1)\,\Delta t)}\right). \tag{37}$$

Under the SDE model (4), it follows that the $\{R_i\}$ are independent samples from a Gaussian distribution with mean $\left(a - \frac{1}{2}b^2\right)\Delta t$ and variance $b^2\,\Delta t$. A key step in Black–Scholes option valuation is the estimation of the volatility parameter, b, so we will aim to infer the value of b and, for simplicity, assume that a is known. More precisely, we seek a *posterior distribution*—a density function that quantifies our degree of belief about possible values of b.

Our SDE model allows us to calculate the probability of any data set $\{R_i\}$ arising, given a value for b. Bayes' theorem makes it possible to turn this around and calculate the probability of any particular value of b arising, given a set of observations $\{R_i\}$. The key relationship is

$$P(b|\{R_i\}) \propto P(\{R_i\}|b)P(b). \tag{38}$$

Here,

- $P(b|\{R_i\})$ is the probability of, or *degree of belief in*, the parameter b, given the data $\{R_i\}$. Our aim is to quantify this *posterior probability* over possible values of b, and the right-hand side of (38) makes this feasible.

- $P(\{R_i\}|b)$ is the probability of the data $\{R_i\}$ arising, given the value of the parameter b. This *likelihood* is made available to us by the model. In our case, it has the form $\prod_{i \geqslant 1} p(R_i; (a - b^2/2)\,\Delta t, b^2\,\Delta t)$, where $p(x; \lambda, \mu^2) = \exp(-(x - \lambda)^2/(2\mu^2))/\sqrt{2\pi\,\mu^2}$ is the density for a Gaussian with mean λ and variance μ^2.
- $P(b)$ is the probability or degree of belief that we assign to b before we see the data. Specifying this *prior probability* is an unavoidable requirement in a Bayesian analysis.

In Fig. 10, we illustrate this idea. Rather than take real financial data, we generated synthetic data using the SDE model with $S(0) = 1$, $a = 0.06$ and $b = 0.4$. In this way, we may judge the quality of our inference. The upper picture shows data for one year, that is, 240 working days. We used a prior distribution that is uniform over $(0.2, 0.6)$—so, before seeing the data we took the view that b must be between 0.2 with 0.6 with all values being equally probable. In other words, $P(b)$ in (38) is constant for $0.2 < b < 0.6$ and zero elsewhere. In the lower picture, we show the posterior distribution that arises when we use the first three months (dotted), six months (dashed) and one year (solid) of data. To make the pictures easier to interpret, we have normalized the densities to have maximum value of one, rather than unit area. For this synthetic experiment, we know that the 'correct' value is $b = 0.4$. We see from the figure that as more data are used, the posterior distribution becomes more sharply peaked and begins to focus on this value.

In this very simple setting, the Bayesian picture is very closely related to the more traditional computational mathematics approach of forming a least-squares objective function (analogous to the

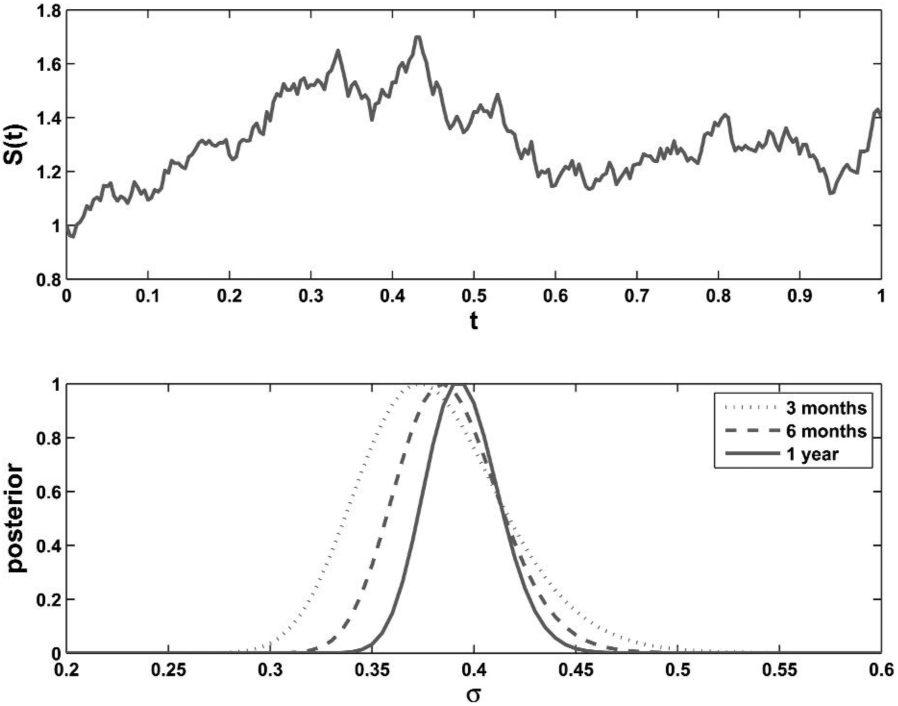

FIG. 10. Upper: one year of asset data from an SDE model. Lower: posterior distribution of the volatility based on three months, six months and one year of data, using a uniform prior distribution.

log-likelihood), adding a penalty function (analogous to the log of the prior) and optimizing to find a single best parameter value (analogous to computing a point that maximizes the posterior). However, working in terms of the complete posterior density, rather than just presenting an optimal parameter and possibly computing local sensitivity around that value, has benefits when there is more than one region of likely values. Also, by sampling parameter values from the posterior, we can display a set of ‘likely’ trajectories from the model.

A further advantage of the Bayesian approach is that higher levels of inference can be performed. If there are two or more plausible models, then there is a systematic framework for simultaneously calibrating and comparing them, even when the models have different numbers of parameters. *Model selection* computations of this nature have been performed on realistic ODE models in systems biology (Vyshemirsky & Girolami, 2008) and much less realistic ODE models in science fiction (Calderhead *et al.*, 2011), and, in principle could be used in the SDE setting.

Many challenges must be overcome if Bayesian inference and model selection are to become mainstream activities in the SDE context. Perhaps, the biggest hurdle is high dimensionality. If N parameters are to be inferred then the posterior distribution is a scalar-valued function of N variables. Searching through $\mathbb{R}^N$ in order to find regions where the posterior takes on significant values is, in general, a huge task—in many inference contexts, this task is more challenging than deterministic global optimization over $\mathbb{R}^N$ in the sense that *all* regions of significant behaviour are required since we must (a) normalize the posterior to have unit area and (b) integrate the posterior across several of the dimensions. Hand in hand with the computational complexity, there is also a visualization issue. How do we display a 25-dimensional random variable to our colleagues? 1D or 2D slices through the posterior, or marginals, where all but 1D or 2D have been integrated out, can be useful, but they necessarily compress information—for example, the globally most likely parameter set according to the full posterior may be very different from the locations of the peaks in these lower dimensional analogues. More fundamentally, unlike the simple example in Fig. 10, in general, the SDE model will not have a known solution, and hence discretization methods will be required in order to construct an approximate likelihood. Added complexity arises if the data itself is assumed to be in error, perhaps in a manner that is correlated across time.

There are many other approaches to SDE model calibration, for example, the recent text (Iacus, 2008) gives examples, many of them *ad hoc* and based on the particular form of the problem, with an emphasis on mathematical finance. This seems to be an area where a general set of principles, bringing together ideas from statistics, applied mathematics and computer science, is yet to emerge. To emphasize that there are possible pitfalls for the unwary, let us return to the asset data example, and suppose that we wish to infer the mean of our log-returns. The intuitively appealing sample average

$$\frac{1}{M}\sum_{i=1}^{M} R_i$$

has the unfortunate property of telescoping down to

$$\frac{1}{M}\sum_{i=1}^{M} \log S(i\,\Delta t) - \log S((i-1)\,\Delta t) = \frac{1}{M}\sum_{i=1}^{M} \log\left(\frac{S(M\,\Delta t)}{S(0)}\right).$$

Hence, this quantity involves *only the first and last observation*, ignoring the vast majority of the data!

In discussing parameter inference in an SDE model, we came up against the task of sampling from the density of a high-dimensional random variable. This general problem arises in many settings (Robert

& Casella, 2004), and it is interesting to note that SDEs, and their numerical discretizations, can play a role. A very useful example of a Markov chain Monte Carlo method known as the *Metropolis-adjusted Langevin algorithm* computes samples from $\pi : \mathbb{R}^N \to \mathbb{R}$ using the SDE

$$\mathrm{d}X(t) = \nabla \log \pi(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W(t).$$

Discretizing this SDE over a long time interval is one approach to sampling from π, and the bias can be eliminated with a suitable acceptance/rejection strategy. There are many practical and theoretical issues to be addressed, including the optimal choice of timestep (Beskos *et al.*, 2009) and level of implicitness (Beskos *et al.*, 2008) in the numerical method, and the development of customized stochastic integrators that preserve geometric structures (Girolami & Calderhead, 2011).

8. Outlook

Overall, this overview of the use of SDEs in applied mathematics, which is naturally biased towards the author's knowledge base and interests, has emphasized five main themes where future activity is likely to have a high impact.

Theoretical issues regarding existence and uniqueness of solutions for non-linear problems, and corresponding results on convergence, stability and the preservation of qualitative features for numerical simulation.

The role of SDEs in multiscale modelling scenarios, especially in systems biology, which will require new theory and tools for hybrid discrete/real-valued models.

More effective Monte Carlo computations in the SDE setting through the use of multilevel methods.

General purpose inference and model selection techniques for quantifying uncertainty.

The use of SDEs and their customized discretizations in a Markov chain Monte Carlo setting to compute samples from a target distribution, typically within a parameter estimation or model calibration exercise.

References

Ait-Sahalia, Y. (1999) Testing continuous-time models of the spot interest rate. *Rev. Financ. Stud.*, **9**, 385–426.

Anderson, D. F., Ganguly, A. & Kurtz, T. G. (2011) Error analysis of tau-leap simulation methods. *Ann. Appl. Probab.* (to appear).

Anderson, D. F. & Mattingly, J. C. (2011) A weak trapezoidal method for a class of stochastic differential equations. *Commun. Math. Sci.*, **9**, 301–318.

Avikainen, R. (2009) Convergence rates for approximations of functionals of SDEs. *Finance Stochastics,* **13**, 381–401.

Ball, K., Kurtz, T. G., Popovic, L. & Rempala, G. (2006) Asymptotic analysis of multiscale approximations to reaction networks. *Ann. Appl. Probab.*, **16**, 1925–1961.

Beskos, A., Roberts, G. O. & Stuart, A. M. (2009) Optimal scalings of Metropolis-Hastings algorithms for non-product targets in high dimensions. *Ann. Appl. Probab.*, **19**, 863–898.

Beskos, A., Roberts, G. O., Stuart, A. M. & Voss, J. (2008) MCMC methods for diffusion bridges. *Stochastics Dyn.*, **8**, 319–350.

CALDERHEAD, B., GIROLAMI, M. & HIGHAM, D. J. (2011) Is it safe to go out yet? Statistical inference in a zombie outbreak model. *Academics on Zombies* (Robert Smith? ed.). University of Ottawa Press (to appear).

CAO, Y., GILLESPIE, D. T. & PETZOLD, L. (2005) The slow-scale stochastic simulation algorithm. *J. Chem. Phys.*, **122**, 014116.

COX, D. & MILLER, H. (1965) *The Theory of Stochastic Processes*. London: Methuen.

CYGANOWSKI, S., KLOEDEN, P. & OMBACH, J. (2002) *From Elementary Probability to Stochastic Differential Equations with MAPLE*. Berlin: Springer.

E, W., LIU, D. & VANDEN-EIJNDEN, E. (2005) Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales. *J. Chem. Phys.*, **123**, 194107.

ERBAN, R., KEVREKIDIS, I. G., ADALSTEINSSON, D. & ELSTON, T. C. (2006) Gene regulatory networks: a coarse-grained equation-free approach to multiscale computation. *J. Chem. Phys.*, **124**, 084106.

GADGIL, C., LEE, C. H. & OTHMER, H. G. (2005) A stochastic analysis of first-order reaction networks. *Bull. Math. Biol.*, **67**, 901–946.

GILES, M. B. (2007) Improved multilevel Monte Carlo convergence using the Milstein scheme. *Monte Carlo and Quasi-Monte Carlo Methods.* Springer, pp. 343–358. Available at http://www.springerlink.com/content/x03273ju277j7524/.

GILES, M. B. (2008) Multilevel Monte Carlo path simulation. *Oper. Res.*, **56**, 607–617.

GILES, M. B., HIGHAM, D. J. & MAO, X. (2009) Analysing multi-level Monte Carlo for options with non-globally Lipschitz payoff. *Financ. Stochastics*, **13**, 403–413.

GILES, M. B. & WATERHOUSE, B. J. (2009) Multilevel Quasi-Monte Carlo Path Simulation. *Radon Series Comp. Appl. Math.*, **8**, 165–181.

GILLESPIE, D. T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, **22**, 403–434.

GILLESPIE, D. T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.

GILLESPIE, D. T. (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, **115**, 1716–1733.

GIROLAMI, M. & CALDERHEAD, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods (with discussion). *J. R. Stat. Soc. B*, **73**, 1–37.

HAIRER, E. & WANNER, G. (1996) *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, 2nd edn. Berlin: Springer.

HASTY, J., PRADINES, J., DOLNIK, M. & COLLINS, J. J. (2000) Noise-based switches and amplifiers for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 2075–2080.

HIGHAM, D. J. (2000) Mean-square and asymptotic stability of the stochastic theta method. *SIAM J. Numer. Anal.*, **38**, 753–769.

HIGHAM, D. J. (2004) *An Introduction to Financial Option Valuation: Mathematics, Stochastics and Computation*. Cambridge: Cambridge University Press.

HIGHAM, D. J., MAO, X. & STUART, A. M. (2002) Strong convergence of Euler-type methods for nonlinear stochastic differential equations. *SIAM J. Numer. Anal.*, **40**, 1041–1063.

HIGHAM, D. J., MAO, X. & YUAN, C. (2007) Almost sure and moment exponential stability in the numerical simulation of stochastic differential equations. *SIAM J. Numer. Anal.*, **45**, 592–609.

HUTZENTHALER, M., JENTZEN, A. & KLOEDEN, P. E. (2011) Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients. *Proc. R. Soc. A* Available at http://rspa.royalsocietypublishing.org/content/early/2010/12/08/rspa.2010.0348.full.

IACUS, S. M. (2008) *Simulation and Inference for Stochastic Differential Equations: With R Examples*. New York: Springer.

INTEP, S., HIGHAM, D. J. & MAO, X. (2009) Switching and diffusion models for gene regulation networks. *Multiscale Model. Simul.*, **8**, 30–45.

JAHNKE, T. (2010) An adaptive wavelet method for the chemical master equation. *SIAM J. Sci. Comput.*, **31**, 4373–4394.

JAYNES, E. (2003) *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

KLOEDEN, P. & NEUENKIRCH, A. (2007) The pathwise convergence of approximation schemes for stochastic differential equations. *Lond. Math. Soc.*, **10**, 235–253.

KLOEDEN, P. E. & PLATEN, E. (1999) *Numerical Solution of Stochastic Differential Equations*, 3rd printing. Berlin: Springer.

KURTZ, T. G. (1981) *Approximation of Population Processes*. Philadelphia, UA: SIAM.

LEIER, A., MARQUEZ-LAGO, T. & BURRAGE, K. (2008) Generalized binomial tau-leap method for biochemical processes incorporating both delay and intrinsic noise. *J. Chem. Phys.*, **128**, 205107.

LEIMKUHLER, B. & REICH, S. (2005) *Simulating Hamiltonian Dynamics*. Cambridge: Cambridge University Press.

MAO, X. (2007) *Stochastic Differential Equations and Applications*, 2nd edn. Chichester, UK: Horwood.

MATTINGLY, J., STUART, A. M. & HIGHAM, D. (2002) Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, **101**, 185–232.

MATTINGLY, J., STUART, A. M. & TRETYAKOV, M. V. (2010) Convergence of numerical time-averaging and stationary measures via the poisson equation. *SIAM J. Numer. Anal.*, **48**, 552–577.

MIKOSCH, T. (1998) *Elementary Stochastic Calculus (with Finance in View)*. Singapore: World Scientific.

MILSTEIN, G. N. & TRETYAKOV, M. V. (2004) *Stochastic Numerics for Mathematical Physics*. Berlin: Springer.

MILSTEIN, G. N. & TRETYAKOV, M. V. (2005) Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients. *SIAM J. Numer. Anal.*, **43**, 1139–1154.

RENSHAW, E. (1991) *Modelling Biological Populations in Space and Time*. Cambridge: Cambridge University Press.

RIPLEY, B. D. (1987) *Stochastic Simulation*. Chichester: Wiley.

ROBERT, C. P. & CASELLA, G. (2004) *Monte Carlo Statistical Methods*, 2nd edn. Berlin, Germany: Springer.

SIVIA, D. & SKILLING, J. (2006) *Data Analysis: A Bayesian Tutorial*. Oxford: Oxford University Press.

STUART, A. M. (2010) Inverse problems: a Bayesian perspective. *Acta Numerica*, **19**, 451–559.

SZPRUCH, L. & HIGHAM, D. J. (2010) Comparing hitting time behavior of Markov jump processes and their diffusion approximations. *Multiscale Model. Simul.*, **8**, 605–621.

SZPRUCH, L., MAO, X., HIGHAM, D. J. & PAN, J. (2011) Numerical simulation of a strongly nonlinear Ait-Sahalia-type interest rate model. *BIT Numer. Math.* Available at http://www.springerlink.com/content/424tt0076k38685u/.

TALAY, D. (1999) Approximation of invariant measures of nonlinear Hamiltonian and dissipative stochastic differential equations. *Progress in Stochastic Structural Dynamics* (R. Bouc & C. Soize eds), vol. 152. Publication du L.M.A.-C.N.R.S., pp. 139–169.

TALAY, D. & TUBARO, L. (1990) Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, **8**, 94–120.

VYSHEMIRSKY, V. & GIROLAMI, M. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833–839.

WILKIE, J. & WONG, Y. M. (2008) Positivity preserving chemical Langevin equations. *Chem. Phys.*, **353**, 132–138.

WILKINSON, D. J. (2006) *Stochastic Modelling for Systems Biology*. Boca Raton, FL: Chapman & Hall/CRC.

YOUNG, W. M. & ELCOCK, E. W. (1966) Monte Carlo studies of vacancy migration in binary ordered alloys: I. *Proc. Phys. Soc.*, **89**, 735–746.