Comparing Hitting Time Behaviour of Markov Jump Processes and their Diffusion Approximations

Lukasz Szpruch* Desmond J. Higham†

November, 2009

Abstract

Markov jump processes can provide accurate models in many applications, notably chemical and biochemical kinetics, and population dynamics. Stochastic differential equations offer a computationally efficient way to approximate these processes. It is therefore of interest to establish results that shed light on the extent to which the jump and diffusion models agree. In this work we focus on mean hitting time behaviour in a thermodynamic limit. We study three simple types of reaction where analytical results can be derived, and we find that the match between mean hitting time behaviour of the two models is vastly different in each case. In particular, for a degradation reaction we find that the relative discrepancy decays extremely slowly; namely, as the inverse of the logarithm of the system size. After giving some further computational results, we conclude by pointing out that studying hitting times allows the Markov jump and stochastic differential equation regimes to be compared in a manner that avoids pitfalls that may invalidate other approaches.

Key words: *birth and death process, Chemical Langevin Equation, finite difference method, Gillespie Algorithm, mean exit time, square root process, stochastic differential equation, thermodynamic limit.*

2000 Mathematics Subject Classification: 60H10, 65J15

*Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, U.K. E-mail: lukas"at"stams.strath.ac.uk

†Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, UK. Supported by EPSRC grant GR/S62383/01.

1 Introduction

Continuous-time, discrete space Markov jump processes are widely studied as models in the natural sciences [31], especially in population biology [28]. They have also found considerable use in cell biology [4, 17, 18, 22, 32], where a chemical kinetics framework [8, 9, 10, 11, 19, 24] has been adopted. We will use the chemical kinetics terminology and refer to the discrete space Markov process as representing the Chemical Master Equation (CME) modeling regime.

When the population size is sufficiently large, it is reasonable to move upscale from the CME to a stochastic differential equation (SDE). Gillespie [11] shows how this SDE, which we refer to as the Chemical Langevin equation (CLE), can be derived under appropriate modeling assumptions, and more rigorous justification of this type of diffusion limit can also be found [23]. Related issues are also addressed in [13].

Replacing the CME with the CLE typically makes both analysis and simulation more tractable, and the development of multi-scale algorithms that combine elements of both regimes is a highly active research area [2, 3, 4, 5, 30]. It is therefore extremely useful to obtain insights into how accurately the CLE approximates the CME. In the case of first order networks, where reaction rates are linear, the moments of the CME satisfy a closed system of ODEs [6]. Gillespie [10] has shown that for first order networks involving a single species, the CLE correctly reproduces first and second moments of the CLE, but not higher moments in general. This result was generalized in [18], where it was shown that the CLE preserves first and second moments and correlations for any first order network.

Although means, variances and correlations are clearly important, there are, of course, many other senses in which we may judge the ability of the continuous-valued CLE to approximate the discrete-valued CME. In this work we look at hitting times—how long does it take a population size to reach a specified upper and/or lower bound? Hitting times arise naturally in many stochastic modeling scenarios. For example,

- in a bi-stable gene regulation system, how long will it take to switch between states [33]?
- in an integrate-and-fire model, when is the next firing [27, 29]?
- in mathematical physics, when will a particle cross a potential barrier [7]?

In order to obtain useful analytical expressions, we found it necessary to focus on simple first order reactions involving a single species. In this case, hitting

time questions for the CLE can be studied via first-order boundary-value ODEs and corresponding analysis for the CME reduces to problems in linear algebra or sequences and series. In sections 2 and 3 we introduce some background material needed for the analysis, and then in sections 4, 5 and 6 we study the three basic reaction types. In the final section we emphasize how the presence of square roots in the CLE can lead to fundamental issues in analysis, and how focusing on hitting time behaviour avoids possible pitfalls.

2 Background and Notation

We are concerned here with the case of a single species that is involved in one or more reaction. Each reaction, for $1 \leq j \leq L$, is specified in terms of

- the *stoichiometric vector*, which in our case is a scalar $\nu_j \in \mathbb{R}$ taking the value -1, 0 or 1, and
- the *propensity function*, $a_j(x)$.

If we let $X(t)$ denote the number of molecules present at time t, then the stoichiometric vector is defined so that the effect of the jth reaction is to update the state from $X(t)$ to $X(t) + \nu_j$. Here, $\nu_j = -1$ if the jth reaction uses up a molecule and $\nu_j = 1$ if it creates one. The propensity function has the property that the probability of this reaction taking place in the infinitesimal time interval $[t, t + dt)$ is given by $a_j(X(t))dt$. With this set-up, $X(t)$ becomes an integer-valued, continuous-time stochastic process. If we let $p_i(t)$ denote the probability $\mathbb{P}(X(t) = i)$, the process may be characterized through the CME, which is the set of ODEs

$$\frac{dp_i(t)}{dt} = \sum_{j=1}^{L} \left(a_j(i - \nu_j) p_{i-\nu_j}(t) - a_j(i) p_i(t) \right), \qquad \text{for } j = 0, 1, 2, \ldots. \tag{1}$$

The CLE is then defined according to the Ito SDE

$$dy(t) = \sum_{j=1}^{L} \nu_j a_j(y(t)) dt + \sum_{j=1}^{L} \nu_j \sqrt{a_j(y(t))} dW_j(t), \tag{2}$$

where the $W_j(t)$ are independent Brownian motions. Here, at each time t, the concentration is represented by the real-valued random variable $y(t)$.

See, for example, [17, 30] for details of how the CME and CLE are defined for general chemical kinetics systems.

Gillespie [11] showed how the CLE arises naturally as an approximation to the CME under certain modeling assumptions; in particular, the number of molecules in the system is required to be "sufficiently large". This is intuitively clear, as the CLE uses a real-valued concentration rather than an integer-valued molecule count. Consequently, in comparing the two processes, we would anticipate that the match should be better for large numbers of molecules. We will give results that formalize this principle in the context of hitting times.

3 Hitting Times for Markov Jump and Diffusion Process

3.1 Hitting time for a Markov Jump Process

At the CME level, the scalar process that we study may be regarded as a Markov jump, or *birth and death* process. Introducing discrete states $\{0, 1, 2, \ldots, M\}$ we may regard states 0 and M as *absorbing*, and define a birth and death process $Z(t)$ with birth rates $\{b_i\}_{i=1}^{M-1}$ and death rates $\{d_i\}_{i=1}^{M-1}$. This process may be characterized via the infinitesimal description

$$\begin{aligned} P\left(Z(t+h) = i+1 \mid Z(t) = i\right) &= b_i h + o(h), \\ P\left(Z(t+h) = i-1 \mid Z(t) = i\right) &= d_i h + o(h), \\ P\left(Z(t+h) = i \mid Z(t) = i\right) &= 1 - (b_i + d_i)h + o(h), \end{aligned}$$

with $b_0 = d_0 = b_M = d_M = 0$.

It then follows, see for example, [14], that the linear system

$$\begin{bmatrix} (b_1+d_1) & -b_1 & & & & \\ -d_2 & (b_2+d_2) & -b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -b_{M-2} \\ & & & & -d_{M-1} & (b_{M-1}+d_{M-1}) \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ \vdots \\ U_{M-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix} \tag{3}$$

has a solution $U \in \mathbb{R}^{M-1}$ such that U_j gives the expected time for the process to be absorbed into state 0 or M given a starting state $Z(0) = j$.

In our CME framework (1) with a single species present, $X(t)$ records the number of molecules at time t. We will be concerned with the expected time for the

molecule count to reach either an upper level b or a lower level a, starting from an initial number x of molecules. Consequently, the general state i of the absorbing birth and death process $Z(t)$ is identified with a molecule count of $a+i$ for $X(t)$, and we have $b-a=M$. To avoid confusion when we consider the analogous hitting times for the process $y(t)$ from the CLE, we introduce the notation

$$T_a^X(x) = \inf\{t \geq 0 \text{ such that } X(t) = a, \text{ given } X(0) = x\},$$

and for the pair $a < b$,

$$T_{a,b}^X(x) = T_a^X(x) \wedge T_b^X(x).$$

We then note that $\mathbb{E}\left[T_{a,b}^X(x)\right]$ for $X(0) = x = a+i$ can be found from U_i in (3).

3.2 Hitting time for Diffusion

In this section we summarize some existing theory concerning hitting times for an SDE. For further details, we refer to [7, 21]. For convenience, we introduce general drift and diffusion coefficients, $\mu(x)$ and $\sigma(x)$, and consider a general scalar Ito SDE with a single Brownian motion

$$dy(t) = \mu(y)dt + \sigma(y)\, dW(t).$$

When they exist, we may then define the *scale function*

$$S(x) = \int^x s(l)dl, \tag{4}$$

where

$$s(x) = \exp\left(-\int^x \frac{2\mu(l)}{\sigma^2(l)}dl\right),$$

and the *speed measure*

$$m(x) = \frac{1}{\sigma^2(x)s(x)}. \tag{5}$$

We introduce the hitting time for the point a and the pair $a < b$ as

$$T_a^y(x) = \inf\{t \geq 0 \text{ such that } y(t) = a, \text{ given } y(0) = x\}$$

and

$$T_{a,b}^y(x) = T_a^y(x) \wedge T_b^y(x).$$

Next, we define the operator L by

$$LV = \mu(x)\frac{dV}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2V}{dx^2}, \qquad \text{for } \ a < x < b.$$

Given a fixed initial condition $x \in (a, b)$, the probability that $y(t)$ hits b before a has the characterization

$$\mathbb{P}\left(T_b^y(x) < T_a^y(x)\right) = \frac{S(x) - S(a)}{S(b) - S(a)}.$$

Similarly,

$$\mathbb{P}\left(T_a^y(x) < T_b^y(x)\right) = \frac{S(b) - S(x)}{S(b) - S(a)}. \tag{6}$$

Next we let u and w denote the solutions to the two-point boundary value ODEs

$$Lu = 0, \quad \text{for} \;\; a < x < b, \qquad u(a) = 0, \quad u(b) = 1,$$

and

$$Lw = -1, \quad \text{for} \;\; a < x < b, \qquad w(a) = 0, \quad w(b) = 0. \tag{7}$$

It follows that $w(x)$ characterizes a mean hitting time,

$$w(x) = \mathbb{E}\left[T_{a,b}^y(x)\right] \tag{8}$$

and we also have

$$\begin{aligned} w(x) \;\; &= \;\; 2\{u(x) \int_x^b [S(b) - S(l)] m(l) dl \\ &\quad + [1 - u(x)] \int_a^x [S(l) - S(a)] m(l) dl\}. \end{aligned} \tag{9}$$

Finally, we introduce some definitions relating to boundary behaviour.

Definition 3.1. *The boundary l is* attracting *if $S(x_0) - S(l) < \infty$ for any $x_0 \in (l, r)$.*

Definition 3.2. *Letting*

$$\Sigma(l) = \lim_{a \searrow l} \int_b^x \left[S(\xi) - S(a)\right] m(\xi) d\xi, \tag{10}$$

the boundary l is said to be attainable *if $\Sigma(l) < \infty$, and* unattainable *if $\Sigma(l) = \infty$.*

Definition 3.3. *Letting*

$$M(l, x] = \lim_{a \searrow l} M[a, x] = \lim_{a \searrow l} \int_a^x m(s) ds, \tag{11}$$

the boundary l is said to be absorbing *if $M(l, x] = \infty$, and* reflective *if $M(l, x] = 0$.*

3.3 Finite Difference Connection

The linear system (3) may be written in the form

$$\frac{b_i + d_i}{2}\left(U_{i+1} - 2U_i + U_{i-1}\right) + \left(b_i - d_i\right)\frac{U_{i+1} - U_{i-1}}{2} = -1, \quad 1 \le i \le M-1.$$

This corresponds to a standard finite difference approximation to the ODE

$$\frac{b(x) + d(x)}{2} u''(x) + \left(b(x) - d(x)\right) u'(x) = -1, \qquad u(a) = u(b) = 0,$$

using a mesh spacing of $\Delta x = 1$; see, for example, [26]. Studying the connection between the discrete linear system and the continuous ODE is the essence of this work. In a numerical analysis setting, the discrete object is regarded as an approximation to the continuous, whereas in our context the opposite is true. However our objective of examining the difference between the two in the large system size, $M \to \infty$, limit is comparable with traditional convergence theory in numerical analysis. A major challenge, however, is that the interval (a, b) is not fixed, and hence the traditional style of error bound, see, for example, [26, Theorem 6.1.3], is not useful. (Equivalently, if we rescale the ODE to the interval $0 \le x \le 1$, then the ODE itself becomes dependent upon M and the coefficients do not have bounded Lipschitz constants.) Indeed, as we will see in section 6, to obtain a positive result in this context it may be necessary to measure the error relative to the (growing) solution size as $M \to \infty$. Related issues arose in the small world analysis of [15, 16].

For this reason, we content ourselves with an investigation of specific simple reactions where the asymptotic behaviour of the discrete and continuous systems can be found and then compared.

4 Production from a source

Perhaps the simplest chemical reaction has the form

$$\emptyset \xrightarrow{k} \mathbf{X} \tag{12}$$

Here members of the species X are being generated at a rate that is independent of the state of the system. (More realistically, X may be generated at a rate proportional to some some other species Y, where Y is sufficiently abundant that its level may be regarded as constant.)

4.1 Discrete Process

From a population dynamics viewpoint, the CME for (12) defines a pure birth process, with population-independent birth rate. An infinitesimal description of this Markov process $X(t)$ is

$$P\left(X(t+h)=s+1 \mid X(t)=s\right)=k\,h+o(h),$$

$$P(X\left(t+h\right)=s-1 \mid X(t)=s)=0,$$

$$P(X\left(t+h\right)=s \mid X(t)=s)=1-k\,h+o(h).$$

In this simple setting, the times between successive births are independent exponentially distributed random variables with mean $1/k$. This has the following immediate consequence, which could also be proved from the general system (3).

Lemma 4.1. *Given integers $b>a\geq 0$, for any integer $x\in(a,b)$ the discrete state Markov process model for (12) with initial molecule count $X(t)=x$ satisfies*

$$\mathbb{E}\left[T^{X}_{a,b}(x)\right]=\frac{b-x}{k}. \tag{13}$$

4.2 Diffusion Process

The CLE for (12) has the form

$$dy(t)=k\,dt+\sqrt{k}\,dW(t). \tag{14}$$

So $y(t)$ is a Brownian motion with drift.

The following lemma characterizes the mean hitting time.

Lemma 4.2. *For any $0\leq a<x<b$, the CLE (14) with initial condition $y(0)=x$ satisfies*

$$\begin{aligned}\mathbb{E}\left[T^{y}_{a,b}(x)\right] &= \frac{1}{k}\left[\frac{-e^{-2x}+e^{-2a}}{-e^{-2b}+e^{-2a}}\left(\frac{-e^{-2b}}{2}\left(e^{2b}-e^{2x}\right)+b-x\right)\right.\\ &\quad\left.+\left(1-\frac{-e^{-2x}+e^{-2a}}{-e^{-2b}+e^{-2a}}\right)\left(a-x+\frac{e^{-2a}}{2}\left(e^{2x}-e^{2a}\right)\right)\right].\end{aligned} \tag{15}$$

Proof. The scale function and speed measure have the form $S(x) = -e^{-2x}/2$ and $m(x) = e^{2x}/k$. This gives

$$\int_x^b [S(b) - S(l)]m(l)dl = \frac{1}{2k}\left(-\frac{e^{-2b}}{2}\left(e^{2b} - e^{2x}\right) + b - x\right)$$

and

$$\int_a^x [S(l) - S(a)]m(l)dl = \frac{1}{2k}\left(a - x + \frac{e^{-2a}}{2}(e^{2x} - e^{2a})\right),$$

and the result follows from (9). □

We note that unlike the discrete process, $X(t)$, the process $y(t)$ in (14) does not preserve positivity. For example, using the expression for $S(x)$ it follows in (6) that

$$\lim_{a\searrow 0}\lim_{b\to\infty} \mathbb{P}\left(T_a^y(x) < T_b^y(x)\right) = \lim_{a\searrow 0}\lim_{b\to\infty} \frac{e^{-2b} - e^{-2x}}{e^{-2b} - e^{-2a}} = e^{-2x}.$$

This limiting probability, however, is small when the initial state is large, in line with our intuition about the relevance of the CLE.

We wish to formalize a sense in which the two hitting times could be close when there is a "large" number of molecules. For this purpose, we will consider an asymptotic regime where the upper limit b tends to infinity and the initial molecule count has the form $x = \alpha b \in \mathbb{Z}$ for some fixed $\alpha \in (0, 1)$. So the initial data scales linearly with the upper exit level. To keep expressions compact we will also set $a = 0$, or, where necessary, take the limit $a \searrow 0$. We will refer to this asymptotic setting as the Large Molecule Count Regime (LMCR).

Theorem 4.1. *In the LMCR the mean hitting times in (13) and (15) satisfy*

$$\left|\mathbb{E}\left[T_{0,b}^X(\alpha b)\right] - \mathbb{E}\left[T_{0,b}^y(\alpha b)\right]\right| \le Ce^{-b\min\{2(1-\alpha),\alpha)\}}, \tag{16}$$

for a constant C independent of b.

Proof. Setting $x = \alpha b \in \mathbb{Z}$ and considering the limit $b \to \infty$, we see in (15) that

$$\begin{aligned}\mathbb{E}\left[T_{a,b}^y(\alpha b)\right] &= \frac{1}{k}\left[\left(1 + O(e^{-2\alpha b})\right)\left(-\tfrac{1}{2} + b(1-\alpha) + O(e^{-2b(1-\alpha)})\right)\right.\\ &\quad \left. + \left(1 + O(e^{-2b(1-\alpha)})\right)\frac{e^{-2\alpha b}}{e^{-2a}}\left(a - \alpha b + \frac{e^{2a}}{2}(e^{2\alpha b} - e^{2a})\right)\right]\end{aligned}$$

and so

$$\mathbb{E}\left[T_{0,b}^y(\alpha b)\right] = \frac{1}{k}\left[-\tfrac{1}{2} + b(1-\alpha) + O(e^{-2b(1-\alpha b)} + e^{-2\alpha b}) + \tfrac{1}{2} + O(be^{-2\alpha b})\right].$$

The result then follows from (13). □

Theorem 4.1 shows that in the LMCR the two hitting times, which grow linearly with b, have an absolute difference that converges exponentially fast. We illustrate this result in Figure 1. Here we took $k = 5$, $a = 0$, $\alpha = \frac{1}{2}$ and chose values of b from 2 to 30 in steps of 2. The hitting time discrepancy on the left-hand side of (16) is seen to decay faster than linearly on a log-log scale.

5 Production

In the production reaction

$$X \xrightarrow{c} 2\,X, \tag{17}$$

new individuals are created at a rate proportional to the current state.

5.1 Discrete Process

In the CME setting, an infinitesimal description for the reaction (17) is

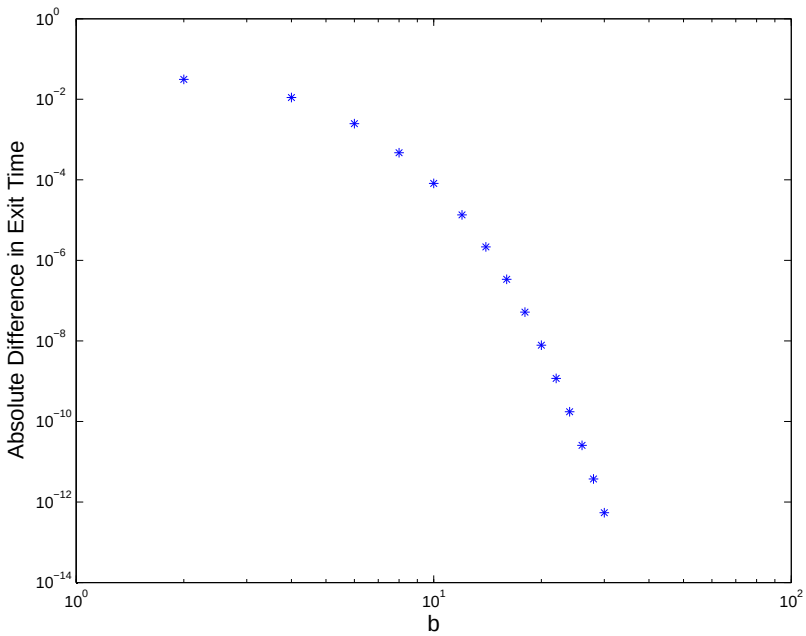

Figure 1: The difference between CME and CLE hitting times for production from a source (12) in the large molecule count regime of Theorem 4.1, on a log-log scale. Horizontal axis is b and vertical axis is the absolute difference on the left-hand side of (16).

$$P(X(t+h) = s+1 \mid X(t) = s) = c\,s\,h + o(h),$$

$$P(X(t+h) = s-1 \mid X(t) = s) = 0,$$

$$P(X(t+h) = s \mid X(t) = s) = 1 - c\,s\,h + o(h).$$

In population dynamics, this corresponds to a pure birth or Yule process [28]. The following result can be obtained from (3) or by simply observing that the times between entering and leaving state s are independently and exponentially distributed random variables with mean $1/(c\,s)$.

Lemma 5.1. *Given integers $b > a \geq 0$, for any integer $x \in (a, b)$ the discrete state Markov process model for (17) with initial molecule count $X(t) = x$ satisfies*

$$\mathbb{E}\left[T^X_{a,b}(x)\right] = \frac{1}{c}\sum_{s=x}^{b-1}\frac{1}{s}. \tag{18}$$

5.2 Diffusion Process

The CLE for (17) has the form

$$dy(t) = c\,y(t)\,dt + \sqrt{c\,y(t)}\,dW(t). \tag{19}$$

This is an example of a mean-reverting square root processes, and a unique non-negative solution is guaranteed [25]. The hitting time may be characterised as follows.

Lemma 5.2. *For any $0 < a < x < b$, the CLE (19) with initial condition $y(0) = x$ satisfies*

$$\begin{aligned}\mathbb{E}\left[T^y_{a,b}(x)\right] &= \frac{1}{c}\left(\frac{e^{-2x} - e^{-2a}}{e^{-2b} - e^{-2a}}\left(-e^{-2b}\int_x^b \frac{e^{2l}}{l}dl + \ln b - \ln x\right)\right. \\ &\quad \left. + \left[1 - \frac{e^{-2x} - e^{-2a}}{e^{-2b} - e^{-2a}}\right]\left(\ln a - \ln x + e^{-2a}\int_a^x \frac{e^{2l}}{l}dl\right)\right), \end{aligned} \tag{20}$$

and for $a = 0$ we may take the limit $\lim \searrow 0$ in this expression.

Proof. The scale function and speed measure have the form $S(x) = -e^{-2x}/2$ and $m(x) = e^{2x}/(cx)$. Hence,

$$\int_x^b [S(b) - S(l)]m(l)dl = \frac{1}{2c}\left(-e^{-2b}\int_x^b \frac{e^{2l}}{l}dl + \ln b - \ln x\right)$$

and

$$\int_a^x [S(l) - S(a)]m(l)dl = \frac{1}{2c}\left(\ln a - \ln x + e^{-2a}\int_a^x \frac{e^{2l}}{l}dl\right),$$

and the result follows from (9). □

We note that neither (18) nor (20) could be considered as closed form expressions. However, both characterizations are amenable to asymptotic analysis, and we may obtain an analogue of Theorem 4.1.

Theorem 5.1. *In the LMCR the mean hitting times in (18) and (20) satisfy*

$$\left|\mathbb{E}\left[T_{0,b}^X(\alpha b)\right] - \lim_{a\searrow 0}\mathbb{E}\left[T_{a,b}^y(\alpha b)\right]\right| \leq Cb^{-2}, \tag{21}$$

for a constant C independent of b.

Proof. First we note that the Harmonic series has the asymptotic expansion

$$\sum_{s=1}^{n}\frac{1}{s} = \ln n + \gamma + \frac{1}{2n} + O(n^{-2}), \qquad \text{as } n \to \infty, \tag{22}$$

where $\gamma = 0.5772\ldots$ is the *Euler-Mascheroni constant* [1].

For the CME we have, using (18) and (22)

$$\begin{aligned}
\mathbb{E}\left[T_{0,b}^X(\alpha b)\right] &= \frac{1}{c}\left[\sum_{s=1}^{b-1}\frac{1}{s} - \sum_{s=1}^{\alpha b-1}\frac{1}{s}\right] \\
&= \frac{1}{c}\left[\ln(b-1) + \gamma + \frac{1}{2(b-1)} - \ln(\alpha b - 1) - \gamma - \frac{1}{2(\alpha b - 1)} + O(b^{-2})\right] \\
&= \frac{1}{c}\left[-\ln\alpha - \frac{1}{2b} + \frac{1}{2\alpha b} + O(b^{-2})\right].
\end{aligned} \tag{23}$$

Next we introduce the exponential integral

$$E_i(x) = \int_{-\infty}^{x}\frac{e^t}{t}\,dt, \quad \text{for } x > 0,$$

and note the asymptotic result

$$E_i(x) = \gamma + \ln x + o(1), \qquad \text{as } x \to 0, \tag{24}$$

see for example, [1]. At the large x extreme, the expansion

$$E_i(x) = \frac{e^x}{x}\left(1 + \frac{1}{x} + O(x^{-2})\right), \qquad \text{as } x \to \infty, \tag{25}$$

follows via integration by parts; see, for example, [20, Chapter 1]. We also have the straightforward identity

$$\int_a^x \frac{e^{2l}}{l}\, dl = E_i(2x) - E_i(2a). \tag{26}$$

Then, using (25) and (26), for any fixed $0 < a \leq a^\star$ we have

$$\begin{aligned}
\frac{e^{-2\alpha b} - e^{-2a}}{e^{-2b} - e^{-2a}} \left(-e^{-2b} \int_{\alpha b}^{b} \frac{e^{2l}}{l} dl + \ln b - \ln \alpha b \right) &= \left(1 + O(e^{-2\alpha b})\right) \times \\
&\quad \left(-\frac{1}{2b} + \ln b - \ln \alpha b + O(b^{-2}) \right) \\
&= -\ln \alpha - \frac{1}{2b} + O(b^{-2}). \qquad (27)
\end{aligned}$$

Similarly, we find that

$$\begin{aligned}
\left(1 - \frac{e^{-2\alpha b} - e^{-2a}}{e^{-2b} - e^{-2a}}\right) \left(\ln a - \ln \alpha b + e^{-2a} \int_a^{\alpha b} \frac{e^{2l}}{l} dl \right) &= \left(e^{-2\alpha b} + O(e^{-2b})\right) \times \\
&\quad \left(\ln a - \ln \alpha b - e^{-2a} \left(E_i(2\alpha b) - E_i(2a)\right)\right).
\end{aligned}$$

Now it follows from (24) that $\ln a - e^{-2a} E_i(2a)$ is bounded for all small a, and hence we find that

$$\begin{aligned}
\left(1 - \frac{e^{-2\alpha b} - e^{-2a}}{e^{-2b} - e^{-2a}}\right) \left(\ln a - \ln \alpha b + e^{-2a} \int_a^{\alpha b} \frac{e^{2l}}{l} dl \right) &= e^{-2\alpha b} E_i(2\alpha b) + O(b^{-2}) \\
&= \frac{1}{2\alpha b} + O(b^{-2}). \qquad (28)
\end{aligned}$$

Combining (20), (23), (27) and (28) gives the required result. □

In Figure 2 we illustrate Theorem 5.1 in the case where $c = 5$, $a = 10^{-3}$ and $\alpha = \frac{1}{2}$ and the upper limit b ranges from 40 to 320 in steps of 40. The hitting time discrepancy on the left-hand side of (21), plotted with asterisks, appears to behave linearly on this log-log scale. A reference line of slope -2 is superimposed. A least squares fit to a power law gave a slope of -2.04 with 2-norm residual of 0.02. This suggests that the $O(b^{-2})$ rate derived in Theorem 5.1 is sharp.

6 Degradation

The degradation reaction may be written

$$X \xrightarrow{c} \emptyset \tag{29}$$

Here a species is undergoing a natural decay process, with a rate that is linearly proportional to the current population size. Intuitively, because of the inherent monotone decrease in the molecule count over time, we would not expect to obtain upper bounds as small as those in Theorems 5.1 and 6.1.

6.1 Discrete Process

In the CME setting, the reaction (29) may be regarded as a pure death process with propensity proportional to $X(t)$. An infinitesimal description is

$$P(X(t+h) = s+1 \mid X(t) = s) = 0,$$

$$P(X(t+h) = s-1 \mid X(t) = s) = c\,s\,h + o(h),$$

$$P(X(t+h) = s \mid X(t) = s) = 1 - c\,s\,h + o(h).$$

As for the case of production, the time between entering and leaving state s is an exponentially distributed random variable with mean $1/(cs)$, and all such times

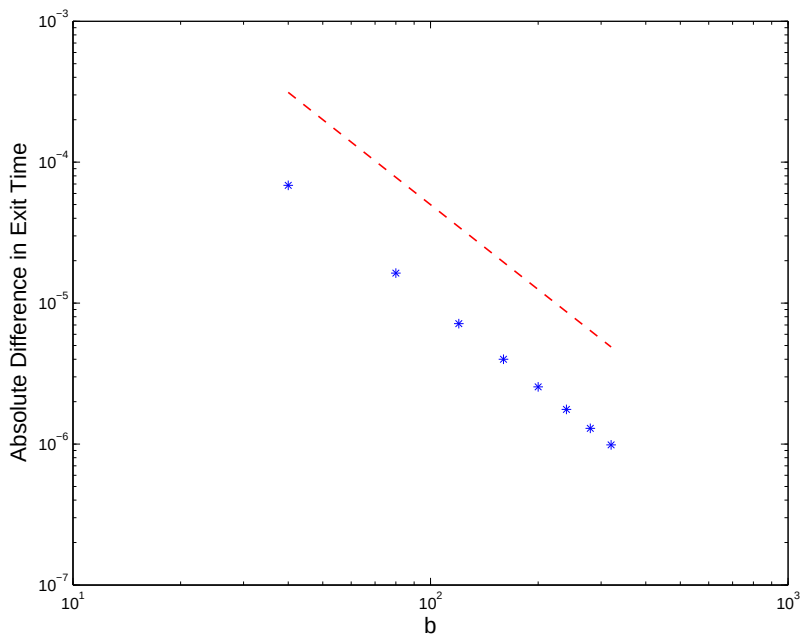

Figure 2: The difference between CME and CLE mean hitting times for a production reaction in the LMC regime of Theorem 5.1, on a log-log scale. Horizontal axis is b and vertical axis is the absolute difference on the left-hand side of (21).

are independent. This leads to the following consequence, which could also be derived from (3).

Lemma 6.1. *Given integers $b > a \geq 0$, for any integer $x \in (a, b)$ the discrete state Markov process model for (29) with initial molecule count $X(t) = x$ satisfies*

$$\mathbb{E}\left[T^X_{a,b}(x)\right] = \frac{1}{c}\sum_{s=a+1}^{x}\frac{1}{s}. \tag{30}$$

6.2 Diffusion Process

The CLE for (29) has the form

$$dy(t) = -cy(t)\,dt - \sqrt{c\,y(t)}\,dW(t). \tag{31}$$

This is another special case from the class of mean-reverting square root processes, and a unique non-negative solution is guaranteed [25]. The hitting time may be characterised as follows.

Lemma 6.2. *For any $0 < a < x < b$, the CLE (31) with initial condition $y(0) = x$ satisfies*

$$\begin{aligned}\mathbb{E}\left[T^y_{a,b}(x)\right] &= \frac{e^{2x}-e^{2a}}{e^{2b}-e^{2a}}\left(\frac{1}{c}\left(e^{2b}\int_x^b \frac{e^{-2l}}{l}dl - \ln b + \ln x\right)\right) \\ &\quad + [1 - \frac{e^{2x}-e^{2a}}{e^{2b}-e^{2a}}]\left(\frac{1}{c}\left(\ln x - \ln a - e^{2a}\int_a^x \frac{e^{-2l}}{l}dl\right)\right). \end{aligned} \tag{32}$$

Proof. The scale function and speed measure have the form $S(x) = e^{2x}/2$ and $m(x) = e^{-2x}/(cx)$. We find that

$$\int_x^b [S(b) - S(l)]m(l)dl = \frac{1}{2c}\left(e^{2b}\int_x^b \frac{e^{-2l}}{l}dl - \ln b + \ln x\right)$$

and

$$\int_a^x [S(l) - S(a)]m(l)dl = \frac{1}{2c}\left(\ln x - \ln a - e^{2a}\int_a^x \frac{e^{-2l}}{l}dl\right),$$

and the result follows from (9). □

We also have

$$\Sigma(0) = \lim_{a\searrow 0}\frac{1}{2c}\left(\ln x - \ln a - e^{2a}\int_a^x \frac{e^{-2l}}{l}dl\right) < \infty$$

and

$$M(0,x] = \lim_{a \searrow 0} \int_a^x m(s)ds = \infty,$$

confirming that zero is an attainable, absorbing boundary.

As in the previous two sections, it is possible to compare mean exit times (30) and (32) in the LMCR to obtain analogues of Theorems 4.1 and 5.1.

Theorem 6.1. *In the LMCR the mean hitting times in (30) and (32) satisfy*

$$\lim_{a \searrow 0} \lim_{b \to \infty} \left(\mathbb{E}\left[T^X_{0,b}(\alpha b)\right] - \mathbb{E}\left[T^y_{a,b}(\alpha b)\right] \right) = \frac{-\ln 2}{c}. \tag{33}$$

Proof. From (30), for the CME as $b \to \infty$ the expansion (22) gives

$$\mathbb{E}\left[T^X_{0,b}(\alpha b)\right] = \frac{1}{c} \sum_{s=1}^{\alpha b} \frac{1}{s} = \frac{1}{c} \left(\ln(\alpha b) + \gamma\right) + o(1). \tag{34}$$

For the CLE, we first consider any fixed value of $a \in (0, a^\star)$, and look at the limit $b \to \infty$. Letting

$$E_1(x) = \int_x^\infty \frac{e^{-t}}{t} dt, \qquad x > 0,$$

denote an alternative type of exponential integral, we may use the asymptotic results

$$E_1(x) = -\ln x - \gamma + o(1), \quad \text{as } x \searrow 0 \tag{35}$$

and

$$E_1(x) = \frac{e^{-x}}{x} + o\left(\frac{e^{-x}}{x}\right), \qquad \text{as } x \to \infty, \tag{36}$$

see, for example, [1, 20]. We also have the identities

$$\int_x^b \frac{e^{-2l}}{l} dl = E_1(2x) - E_1(2b) \quad \text{and} \quad \int_a^x \frac{e^{-2l}}{l} dl = E_1(2a) - E_1(2x). \tag{37}$$

Then using (36) and (37) in the first term on the right-hand side of (32), we have

$$\begin{aligned} \frac{e^{2\alpha b} - e^{2a}}{e^{2b} - e^{2a}} \left(e^{2b} \int_{\alpha b}^b \frac{e^{-2l}}{l} dl - \ln b + \ln \alpha b \right) &= \frac{e^{2\alpha b} - e^{2a}}{1 - e^{2(a-b)}} \left(E_1(2\alpha b) - E_1(2b)\right) \\ &\quad - \frac{e^{2\alpha b} - e^{2a}}{e^{2b} - e^{2a}} \ln \alpha \\ &= o(1), \end{aligned} \tag{38}$$

as $b \to \infty$, uniformly in a.

For the second term on the right-hand side of (32),

$$
\begin{aligned}
[1-\frac{e^{2\alpha b}-e^{2a}}{e^{2b}-e^{2a}}]\left(\ln\alpha b-\ln a-e^{2a}\int_a^{\alpha b}\frac{e^{-2l}}{l}dl\right) &= \left(1+O(e^{-2b(1-\alpha)}\right)(\ln\alpha b-\ln a \\
&\quad -e^{2a}\left(E_1(2a)-E_1(2\alpha b)\right)\right) \\
&= \left(1+O(e^{-2b(1-\alpha)}\right)(\ln\alpha b-\ln a \\
&\quad -e^{2a}E_1(2a)\right). \qquad (39)
\end{aligned}
$$

It follows from (32), (34), (38) and (39) that uniformly in a and for large b,

$$
\mathbb{E}\left[T^X_{0,b}(\alpha b)\right]-\mathbb{E}\left[T^y_{a,b}(\alpha b)\right]=\frac{1}{c}\left(\gamma+\ln a+e^{2a}E_1(2a)+o(1)\right).
$$

Taking the limit $a\searrow 0$ and using (35), we obtain the required result. □

Figure 3 illustrates Theorem 6.1. As for Figure 1 we used $c=5$, $\alpha=\frac{1}{2}$ and chose values of b from 2 to 30 in steps of 2. For the CLE hitting time (32) we took lower limits of $a=10^{-2}$, 10^{-4} and 10^{-8}. We see that as b increases and a decreases, the absolute value of the hitting time discrepancy on the left-hand side of (33) approaches the limiting value $\ln(2)/5\approx 0.1386$ predicted by the theorem.

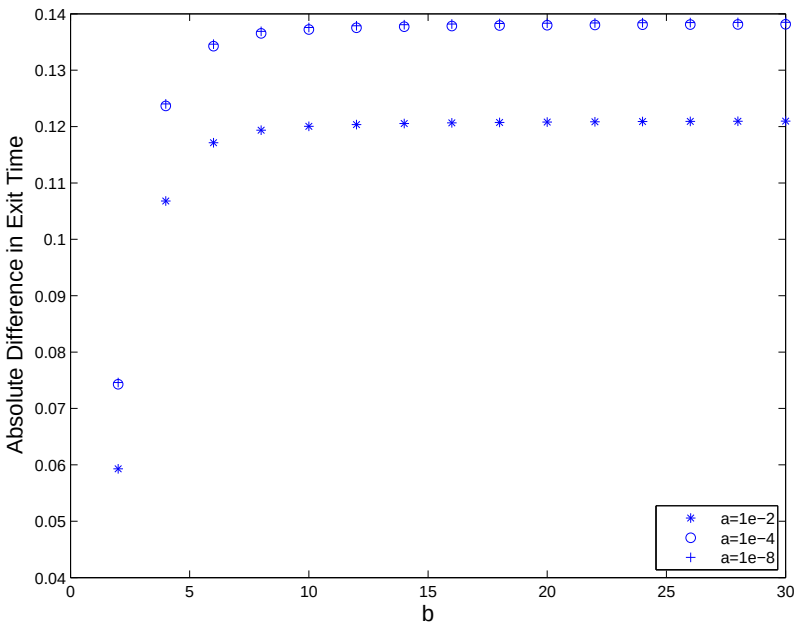

Figure 3: The difference between CME and CLE mean hitting times in the regime of Theorem 5.1. Lower limit of $a=10^{-2}$ asterisks; $a=10^{-4}$ circles; $a=10^{-8}$ plus signs.

There is a stark contrast between the results in Theorems 4.1, 5.1 and 6.1. In the first case, the two mean hitting times converge exponentially quickly in the large molecule count limit, in the second case they converge only at a polynomial rate, and in the third case there is a fixed, nonzero limiting absolute error. Since the two hitting times in Theorem 6.1 grow like $\ln(b)$ as $b \to \infty$, this lack of convergence in an absolute sense translates into convergence in a *relative sense*—the ratio of the mean hitting time discrepancy to the actual mean hitting time tends to zero; albeit at a rate of only $1/\ln b$.

7 Two-way Reactions and General Issues

We begin this final section by briefly discussing difficulties arising with the diffusion regime in the case where pairs of reactions are combined. For production from a source (12) and degradation (29), we obtain the system

$$\emptyset \xrightarrow{k} X \xrightarrow{c} \emptyset \tag{40}$$

Here, the birth and death rates for the CME are $P(X\,(t+h) = s+1 \mid X(t) = s) = k\,h + o(h)$ and $P(X\,(t+h) = s-1 \mid X(t) = s) = c\,s\,h + o(h)$, respectively, and the corresponding CLE has the form

$$dy(t) = (k - c\,y(t))\,dt + \sqrt{k}\,dW_1 - \sqrt{c\,y(t)}\,dW_2, \tag{41}$$

where $W_1(t)$ and $W_2(t)$ are independent scalar Brownian motions.

In this case, where there are two noise sources, the theory in section 3.2 carries through when we interpret $\sigma^2(x)$ in the scale function (4) and speed measure (5) as the sum of the squares of the two diffusion coefficients.

We then find that

$$\begin{aligned}
s(x) &= e^{2x}(k+cx)^{\frac{-4k}{c}}, \\
m(x) &= \frac{\exp(-2x)}{(k+cx)^{1-\frac{4k}{c}}}, \\
S(x) &= 2^{\frac{4k}{c}-1} e^{\frac{-2k}{c}} (k+cx)^{\frac{-4k}{c}} \left(-\frac{k+cx}{c}\right)^{\frac{4k}{c}} \Gamma\left(1-\frac{4k}{c}, -\frac{2\,(k+cx)}{c}\right),
\end{aligned}$$

where Γ denotes the incomplete Gamma function. Thus, even for this relatively simple system, the task of performing an asymptotic LMCR expansion of $w(x)$ in (9) and comparing this with a corresponding asymptotic expansion of the linear system solution in (3) would be extremely daunting, and perhaps intractable.

Applying Definition 3.2, we find that zero is an attainable boundary for the SDE (41). Because the first noise term $\sqrt{k}\,dW_1$ in (41) does not vanish at $y = 0$, the process may then break down, no longer producing a real solution. Hence, a solution to the CLE only makes sense up to a stopping time defined by the solution reaching zero, and any analysis must acknowledge this fact. The hitting time framework deals with this difficulty in a natural manner.

Because direct analysis does not seem possible for (40), Figure 4 reports on a computational test. Here, we took rate constants $k = 1$ and $c = 1$. We used a lower limit of $a = b/4$ for the exit time, and started with $x = b/2$ molecules. Values of $b = 4\times10^1, 4\times10^2, 4\times10^3, 4\times10^4$, were used. As b increases, we are taking a larger system size and, since a scales like b, avoiding the case of small molecule counts. The CME and CLE exit times were computed by solving numerically the sparse linear system (3) and the boundary value ODE (7), respectively. We show the absolute and relative difference between CME and CLE exit times on a log-log scale. In this favourable large molecule regime, we observe convergence of the two hitting times—a least squares fit gives a power of -1.99 with a residual of 0.06, suggesting the same rate of b^{-2} that was rigorously derived in section 5 for a production reaction.

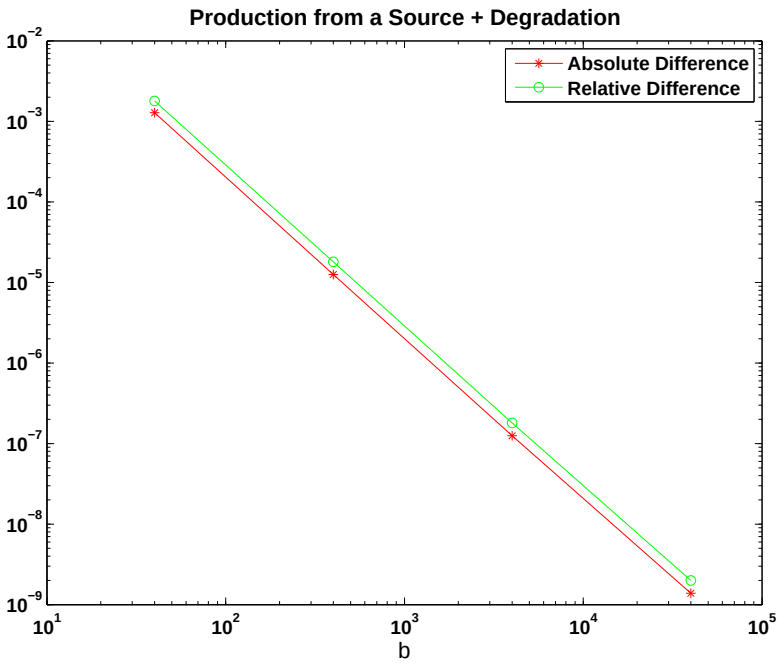

Figure 4: Absolute and relative difference between CME and CLE exit times for production from a source combined with degradation (40), on a log-log scale. Here we have lower limit $a = b/4$ and starting value $x = b/2$.

The failure of the CLE (41) to stay nonnegative is not simply a consequence of the additive noise term. To see this, we may consider the reversible isometry

$$X_1 \underset{c_2\, X_2}{\overset{c_1\, X_1}{\rightleftharpoons}} X_2 \tag{42}$$

Here, a molecule of species X_1 may convert into a molecule of species X_2, with propensity proportional to the number of X_1 molecules, and, similarly, a molecule of species X_2 may convert into a molecule of species X_1, with propensity proportional to the number of X_2 molecules. In this system, the total number of molecules remains constant. Hence, if we assume that there is a deterministic total of K molecules at time zero then we may write a CLE for X_1 alone in the form

$$dy(t) = (-c_1 y(t) + c_2(K - y(t)))\, dt - \sqrt{c_1 y(t)}\, dW_1 + \sqrt{c_2(K - y(t))}\, dW_2. \tag{43}$$

Gillespie [12] uses this example as the basis for comparing steady state distributions for the CME and CLE. In the CLE case, Gillespie claims to solve the steady Fokker-Planck equation, and displays analytical solutions for the resulting distribution. Numerical plots are given to show that the steady CME and CLE distributions are close. However, that CLE analysis is done under the implicit assumption that the stochastic process is well behaved for all time and has a well defined steady distribution. Looking at (43) we see that the noise terms do not both switch off at the 'endpoints': if $y(t) = 0$ then the diffusion coefficient $\sqrt{c_2(K - y(t))}$ is active and if $y(t) = K$ then the diffusion coefficient $-\sqrt{c_1 y(t)}$ is active. Hence, there is no reason to believe that the SDE will remain in the range $[0, K]$. In fact Gillespie's conclusion [12, Eq(15)] that for $c_1 = c_2$ the steady distribution is normal, contains an inherent contradiction—a normal distribution allows a nonzero probability of molecule counts outside the range $[0, K]$, in which case the process is not well behaved. So, while fully agreeing with the comments in [12] that the CLE is inaccurate in the far tails because this is precisely where the modelling assumptions used to derive the CLE become invalid, we wish to make the further point that this invalidity manifests itself even more seriously through a breakdown in the fundamental existence/uniqueness of the stochastic process. Of course, it is possible to 'fix up' the definition of the CLE by introducing reflecting boundary conditions or by taking absolute values inside the square root function, but neither alteration would respect the integrity and elegance of the first principles modelling approach in [11].

In Figure 5 we test a similar scenario to that in Figure 4, this time for the reversible isometry (42). In this case, we used the total number of molecules, K, to control the system size. We took $a = K/8$ and $b = K/2$ for the upper and lower limits, with starting value $x = K/4$. Rate constants were set to $c_1 = c_2 = 1$, and K was varied over $8 \times 10^1, 16 \times 10^1, 8 \times 10^2, 16 \times 10^2, 8 \times 10^3, 16 \times 10^3, 8 \times$

$10^4, 16 \times 10^4, 8 \times 10^5, 16 \times 10^5$. As in Figure 4, by letting the system size increase and avoiding small numbers of molecules, we observe convergence. In this case a least squares fit for the absolute difference gives a power of -0.99 with residual 0.04, suggesting a rate of K^{-1}.

Our tenet here is that a systematic comparison of the CME and CLE regimes must take account of the fact that the square roots in the diffusion coefficients, which arise perfectly naturally through modelling arguments, cause genuine analytical difficulties. These difficulties can be traced back to the modelling assumptions, and they arise when a species becomes scarce. The approach that we take here of comparing the two regimes in terms of hitting times has the benefit of allowing us to focus on the diffusion process before it breaks down, and it gives a realistic way to address the 'thermodynamic limit'.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*

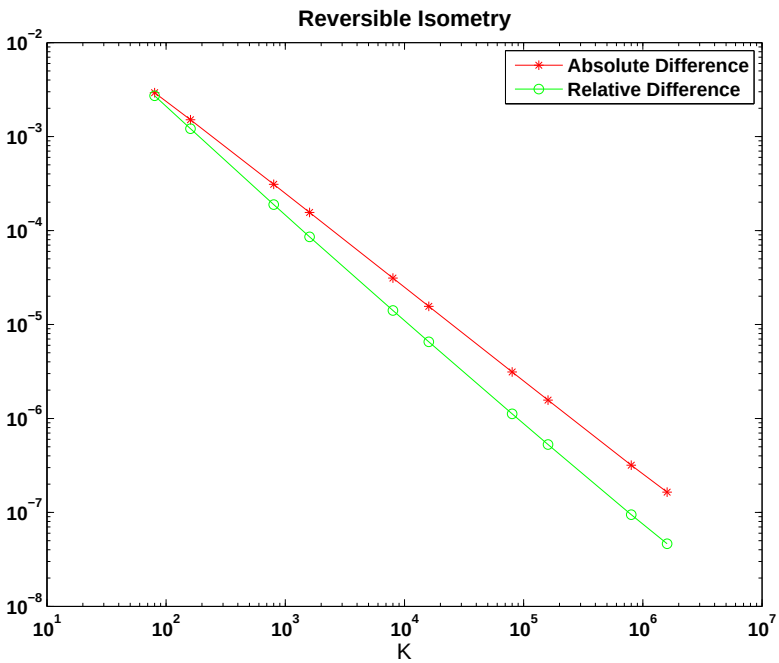

Figure 5: Absolute and relative difference between CME and CLE exit times for the reversible isometry (42), on a log-log scale. Here K is the total number of molecules and we have upper and lower limits $a = K/8$, $b = K/2$ and starting value $x = K/4$.

with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1964.

[2] D. Adalsteinsson, D. McMillen, and T. C. Elston, *Biochemical network stochastic simulator (BioNetS): software for stochastic modeling of biochemical networks*, BMC Bioinformatics, 5:24 (2004).

[3] K. Ball, T. G. Kurtz, L. Popovic, and G. Rempala, *Asymptotic analysis of multiscale approximations to reaction networks*, The Annals of Applied Probability, 16 (2006), pp. 1925–1961.

[4] Y. Cao, D. T. Gillespie, and L. Petzold, *The slow-scale stochastic simulation algorithm*, J. Chem. Phys., 122 (2005), p. 014116.

[5] W. E, D. Liu, and E. Vanden-Eijnden, *Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales*, J. Chem. Phys., 123 (2005), p. 194107.

[6] C. Gadgil, C. H. Lee, and H. G. Othmer, *A stochastic analysis of first-order reaction networks*, Bulletin of Mathematical Biology, 67 (2005), pp. 901–946.

[7] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer, Berlin, 3rd ed., 2004.

[8] D. T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, J. Comp. Phys., 22 (1976), pp. 403—434.

[9] ——, *Exact stochastic simulation of coupled chemical reactions*, J. Phys. Chem., 81 (1977), pp. 2340–2361.

[10] ——, *Markov Processes: An Introduction for Physical Scientists*, Academic Press, San Diego, 1991.

[11] ——, *The chemical Langevin equation*, J. Chem. Phys., 113 (2000), pp. 297–306.

[12] ——, *The chemical Langevin and Fokker-Planck equations for the reversible isomerization reaction*, J. Phys. Chem. A, 106 (2002), pp. 5063–5071.

[13] ——, *Deterministic limit of stochastic chemical kinetics*, J. Phys. Chem. B, 113 (2009), pp. 1640–1644.

[14] J. Glaz, *Probabilities and moments for absorption in finite homogeneous birth-death processes*, Biometrics, 35 (1979), pp. 813–816.

[15] D. J. Higham, *Greedy pathlengths and small world graphs*, Linear Algebra and Its Applications, 416 (2006), pp. 745–758.

[16] ——, *A matrix perturbation view of the small world phenomenon*, SIAM Review, 49 (2007), pp. 91–108.

[17] ——, *Modeling and simulating chemical reactions*, SIAM Review, 50 (2008), pp. 347–368.

[18] D. J. Higham and R. Khanin, *Chemical master versus chemical Langevin for first-order reaction networks*, The Open Applied Mathematics Journal, (2008), pp. 59–79.

[19] S. Intep, D. J. Higham, and X. Mao, *Switching and diffusion models for gene regulation networks*, Multiscale Model. Simul., 8 (2009), pp. 30–45.

[20] D. S. Jones, *Introduction to Asymptotics*, World Scientific, Singapore, 1997.

[21] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, San Diego, 1981.

[22] R. Khanin and D. J. Higham, *Chemical Master Equation and Langevin regimes for a gene transcription model*, Theoretical Computer Science, 408 (2008), pp. 31–40.

[23] T. G. Kurtz, *Approximation of Population Processes*, SIAM, 1981.

[24] T. Li, *Analysis of explicit tau-leaping schemes for simulating chemically reacting systems*, Multiscale Model. Simul., 6 (2007), pp. 417–436.

[25] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, 2007.

[26] J. M. Ortega, *Numerical Analysis: A Second Course*, SIAM, Philadelphia, 1990.

[27] H. E. Plesser, *Noise in integrate-and-fire neurons: From stochastic input to escape rates*, Neural Computation, 12 (2000), pp. 367–384.

[28] E. Renshaw, *Modelling Biological Populations in Space and Time*, Cambridge University Press, 1991.

[29] A. Saarinen, M.-L. Linne, and O. Yli-Harja, *Stochastic differential equation model for cerebellar granule cell excitability*, PLoS Comput Biol, 4 (2008), p. e1000004.

[30] H. E. Samad, M. Khammash, L. Petzold, and D. T. Gillespie, *Stochastic modeling of gene regulatory networks*, Int. J. Robust and Nonlinear Control, 15 (2005), pp. 691–711.

[31] H. M. Taylor and S. Karlin, *An Introduction to Stochastic Modeling*, Academic Press, San Diego, 3rd ed., 1998.

[32] D. J. Wilkinson, *Stochastic Modelling for Systems Biology*, Chapman & Hall/CRC, 2006.

[33] V. P. Zhdanov, *Transient stochastic bistable kinetics of gene transcription during the cellular growth*, Chemical Physics Letters, 424 (2006), pp. 394–398.