Chemical Master Equation and Langevin Regimes for a Gene Transcription Model

Raya Khanin [a] Desmond J. Higham [b]

[a] *Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, U.K.*

[b] *Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK.*

Abstract

Gene transcription models must take account of intrinsic stochasticity. The Chemical Master Equation framework is based on modelling assumptions that are highly appropriate for this context, and the Stochastic Simulation Algorithm (also known as Gillespie's algorithm) allows for practical simulations to be performed. However, for large networks and/or fast reactions, such computations can be prohibitatively expensive. The Chemical Langevin regime replaces the massive ordinary differential equation system with a small stochastic differential equation system that is more amenable to computation. Although the transition from Chemical Master Equation to Chemical Langevin Equation can be justified rigorously in the large system size limit, there is very little guidance available about how closely the two models match for a fixed system. Here, we consider a transcription model from the recent literature and show that it is possible to compare first and second moments in the two stochastic settings. To analyse the Chemical Master Equation we use some recent work of Gadgil, Lee and Othmer, and to analyse the Chemical Langevin Equation we use Ito's Lemma. We find that there is a perfect match—both modelling regimes give the same means, variances and correlations for all components in the system. The model that we analyse involves 'unimolecular reactions', and we finish with some numerical simulations involving dimerization to show that the means and variances in the two regimes can also be close when more general 'bimolecular reactions' are involved.

Key words: chemical kinetics, gene regulation, Gillespie, multi-scale, moments, noise, stochastic simulation, systems biology, transcription, translation.
PACS: 82.20.Wt

1 Background

Several experimental techniques are now available to measure gene expression, even at the single cell level [3,7,17]. In parallel, mathematical models

and simulation algorithms have been developed to explain these observations and make new predictions [1,8,10,24,29,28,27]. Key modeling and simulation challenges in this area are that (a) some components may be present in relatively small quantities, (b) there can be a wide range of natural time scales in operation, and (c) on the level at which observations are made, the process is inherently stochastic. A Markov process, or *Chemical Master Equation* (CME) framework is highly appropriate in this context, and is now widely used. The CME methodology and an accompanying simulation algorithm can be traced back to the work of Gillespie in the chemical kinetics literature [12,13]. Recent overviews can be found in [10,16,19] and we note that there are close connections to Petri nets, discrete event simulation and birth-and-death processes [30].

Because the CME framework takes account of every reaction, for many realistic models it is too computationally expensive to be useful. The *Chemical Langevin Equation* (CLE) provides an alternative model that retains some of the main features of the CME whilst making simulations more feasible. The CLE, which takes the form of an Ito stochastic differential equation (SDE), can be derived from the CME via a series of reasonable modeling assumptions [15,22], and under the extreme case where fluctuations in the CLE are ignored, we recover the traditional deterministic *Reaction Rate Equation* (RRE), or law of mass action. Many authors are now developing *multi-scale* simulation methods that automatically operate in the cheapest modeling regime that captures the appropriate behaviour [6,9]. For this reason it is important to have an understanding of how the different modelling regimes compare. This motivates the work here, where the means and variances of the CME and CLE are compared for a recent gene transcription model. To analyse the CME we make use of the general first-order reaction theory of Gadgil et al. [11] and to analyse the CLE we perform what appears to be the first application of Ito's lemma in this context.

The article is organised as follows. In the next section we give a very simple example that illustrates the main concepts involved in our work. Then in section 3 we set up the general specification of the CME and CLE and introduce Ito's lemma. The gene regulation model is described in section 4 and moments for the CME and CLE are derived analytically in sections 5 and 6 respectively. A numerical experiment involving dimerization is given in section 7 to show that similar behaviour can also arise when we leave the first-order realm.

2 Illustrative Example: Unimolecular Decay

To illustrate the ideas in this work, we begin with the simplest possible type of reaction; unimolecular decay. We suppose that there is only one species, S,

in our system, and the only event that can take place at any time is that one molecule of S may decay. We could write the system symbolically as

$$S \xrightarrow{c} \emptyset. \tag{1}$$

Here, $c > 0$ is a constant that relates to the propensity of the decay process. We suppose that initially, at time $t = 0$, the number of molecules of S is known to be N. The state of the system at time t is fully described by a non-negative integer $X(t)$, representing the number of molecules of S present. So $X(t)$ may take any of the values $N, N-1, N-2, \ldots, 1, 0$. In the CME setting we regard $X(t)$ as a discrete-valued random variable at each point in time, and work in terms of the probability $p_i(t)$ that $X(t) = i$, arriving at the ordinary differential equation (ODE) system

$$\frac{d}{dt}p_N(t) = -cNp_N(t), \tag{2}$$

$$\frac{d}{dt}p_i(t) = c \cdot (i+1) \cdot p_{i+1}(t) - c \cdot i \cdot p_i(t), \quad \text{for } i = N-1, N-2, \ldots, 0. \tag{3}$$

The general ODE (3) has a natural interpretation. The rate of change of $p_i(t)$ has a positive contribution $c \cdot (i+1) \cdot p_{i+1}(t)$, which corresponds to the fact that we arrive at state i via one decay from state $i+1$. Conversely, there is a negative contribution $-c \cdot i \cdot p_i(t)$ due to the fact that, when in state i, we leave that state when a decay takes place.

The system (2)–(3) is readily solved to give

$$p_i(t) = \frac{N!}{i!(N-i)!} e^{-cit} \left(1 - e^{-ct}\right)^{N-i}, \quad \text{for } i = 0, 1, 2, \ldots, N. \tag{4}$$

Using $\mathbb{E}[\cdot]$ and $\mathsf{Var}[\cdot]$ to denote the mean and variance, respectively, it follows that

$$\mathbb{E}\left[X(t)\right] = Ne^{-ct} \quad \text{and} \quad \mathsf{Var}\left[X(t)\right] = Ne^{-ct}\left(1 - e^{-ct}\right). \tag{5}$$

Details can be found, for example, in [26] by observing that this system corresponds to a pure death process in the context of stochastic population modelling.

In the CLE setting, we represent the amount of species S present at time t by the real-valued stochastic process $Y(t)$. In other words, at each time t, $Y(t)$ is a continuous-valued random variable. The CLE is then the Ito SDE [18,23]

$$dY(t) = -cY(t)\,dt - \sqrt{cY(t)}\,dW(t), \quad Y(0) = N. \tag{6}$$

Because the drift coefficient $-cY(t)$ is linear, it follows immediately that $\mathbb{E}\left[Y(t)\right]$ satisfies the ODE that arises when the noise is switched off, giving

$$\mathbb{E}\left[Y(t)\right] = Ne^{-ct}. \tag{7}$$

To find the second moment, we may apply Ito's lemma, as described in section 3.2, to get

$$\frac{d}{dt}\mathbb{E}\left[Y(t)^2\right] = -2\,c\,\mathbb{E}\left[Y(t)^2\right] + c\,\mathbb{E}\left[Y(t)\right].$$

Using the expression (7), this solves to give $\mathbb{E}\left[Y(t)^2\right] = Ne^{-ct} + N(N-1)e^{-2ct}$, so that

$$\mathsf{Var}\left[Y(t)\right] = Ne^{-ct}\left(1 - e^{-ct}\right). \tag{8}$$

Comparing (7) and (8) with (5), we see that the models give precisely the same expressions for the mean and variance of S. This happens despite the fact that one uses the discrete, integer-valued state vector $X(t)$ and the other uses the real-valued $Y(t)$.

For completeness, we mention that the RRE formulation for the system (1) is the scalar ODE $dz(t)/dt = -cz(t)$, where $z(t)$ is a deterministic real-valued quantity representing the amount of S present at time t. This is precisely the ODE for the mean in the CLE, and hence $z(t) = \mathbb{E}[Y(t)] = Ne^{-ct}$.

In Figure 1 the 'descending staircase' formed by the circles indicates a path generated from the CME, computed with Gillespie's *Stochastic Simulation Algorithm* (SSA) [12,13]. Here we fixed the rate constant at $c = 1$ and started with $N = 10$ molecules. We see that the state decreases by one each time there is a "reaction". The jagged curve in a solid line-type shows a path from the CLE, computed with the Euler–Maruyama method [18,23]. The dashed curved shows the solution given by the RRE. We mention that in this special case of pure decay, the state of the underlying physical system is inherently non-increasing with time. All CME paths and the RRE solution reproduce this monotonicity, but a CLE path clearly does not.

For Figure 2 we changed the initial condition to $N = 100$. In this case the differences between the three modelling regimes are much less dramatic; compared to Figure 1, we are starting closer to thermodynamic (large system size) limit where fluctuations become insignificant [2,21].

Two features of the CLE (6) for this simple model are generic.

1 The diffusion coefficient is nonlinear.
2 The description of the problem involves a square root, and hence the problem is only well defined if the solution remains non-negative.

With regard to the second point, the particular CLE (6) is a special case of a *square root process*. These SDEs are widely used as interest rate models in mathematical finance, and it can be shown that the solution in (6) maintains non-negativity with probability one [23]. However, we note that the issue of

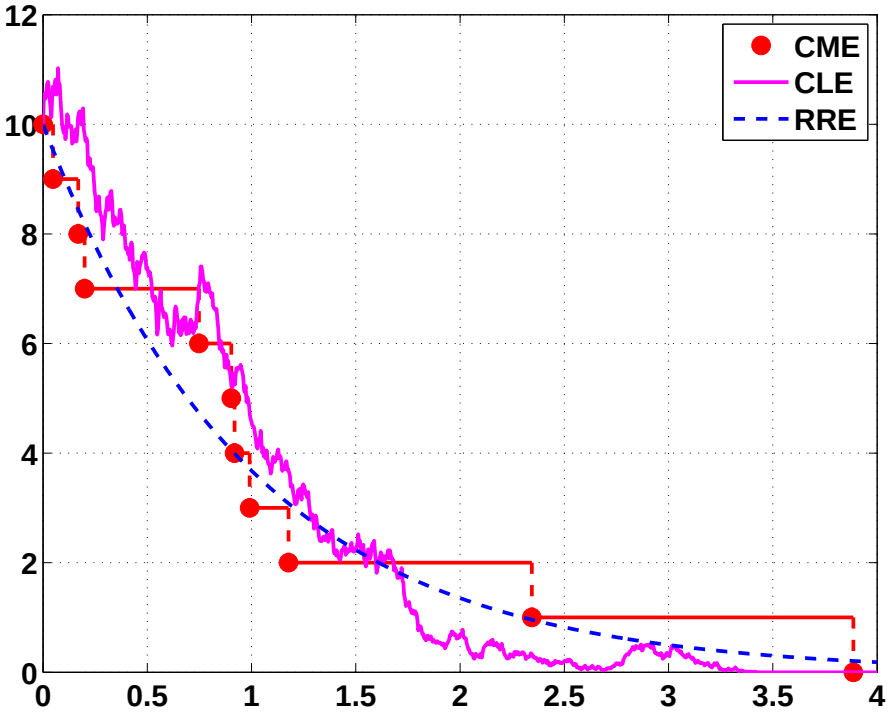

Fig. 1. For the pure decay process (1) with rate constant $c = 1$ and $N = 10$ molecules initially: one path from the CME (circles); one path from the CLE (dashed); and the solution of the RRE (dotted).

negative solutions seems to be open for general CLEs. In this work, we will always assume that the CLE has a well-defined, unique solution.

The main result in this article is that the coincidence of CME and CLE mean and variance in the simple model (1) carries through to a gene transcription model. We note that Gillespie [14, Section 6.2.A] showed the equivalence of first and second (but never third) moments between a birth-death Markov process and the analogous continuous Markov process in the case of scalar model (one species) with linear drift and quadratic diffusion. Our work extends this to the case of a specific system of four species.

3 Stoichiometric Formalization

3.1 Chemical Master Equation

Suppose that there are N chemical species, $S_1, S_2, \ldots, S_N$, taking part in M different chemical reactions. In the CME formulation, we have a state vector $\mathbf{X}(t) \in \mathbb{R}^N$ whose ith component, $X_i(t)$, denotes the number of molecules of S_i present at time t. Hence, each $X_i(t)$ is a non-negative integer. For each

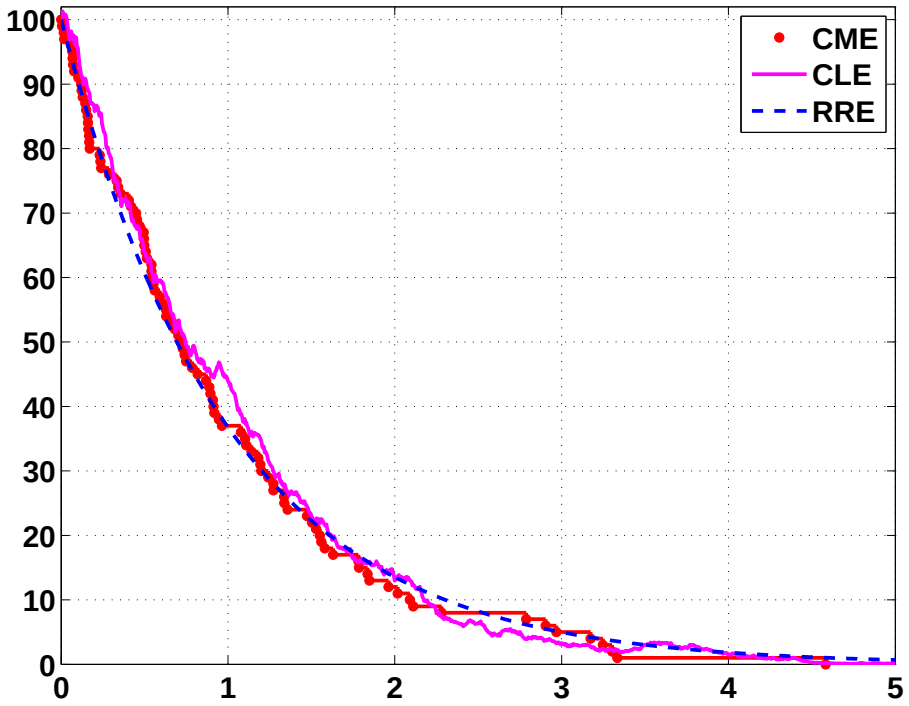

Fig. 2. As for Figure 1 but with $N = 100$ molecules initially.

$1 \leq j \leq M$ we have a *stoichiometric vector* $\boldsymbol{\nu}_j \in \mathbb{R}^N$, and *propensity function* $a_j(\mathbf{X}(t))$, such that the jth reaction takes place over the infinitesimal interval $[t, t+dt)$ with probability $a_j(\mathbf{X}(t))\,dt$ and causes the change $\mathbf{X}(t) \mapsto \mathbf{X}(t)+\boldsymbol{\nu}_j$ to the state vector.

Letting $P(\mathbf{x}, t)$ denote the probability that $\mathbf{X}(t) = \mathbf{x}$, the CME is the ODE system

$$\frac{dP(\mathbf{x},t)}{dt} = \sum_{j=1}^{M} \left(a_j(\mathbf{x}-\boldsymbol{\nu}_j)P(\mathbf{x}-\boldsymbol{\nu}_j,t) - a_j(\mathbf{x})P(\mathbf{x},t)\right). \tag{9}$$

Generally, the CME cannot be solved analytically in any useful way, although, as indicated in Figures 1 and 2, the SSA gives a way to compute realisations of $\{t, \mathbf{X}(t)\}$ that respect the CME. However, in the case where all reactions are unimolecular (or first-order), detailed analysis is possible, both for the first and second moments [11] and the general distributions [20]. In this work we will show that, at least for a specific gene regulation model, useful analytical results can also be derived for the CLE formulation described in the next subsection.

The CLE uses a real-valued random variable $\mathbf{Y}(t) \in \mathbb{R}^N$ to describe the state of the system at time t. The jth component $Y_j(t)$ represents the amount of species j. In moving from the CME to the CLE we (typically) make a dramatic reduction in the number of components, but pay the price that each component is a real-valued random variable, rather than a non-negative integer. The CLE takes the form of an Ito SDE [18,23]

$$d\,\mathbf{Y}(t) = \sum_{j=1}^{M} \boldsymbol{\nu}_j\, a_j(\mathbf{Y}(t))\, dt + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{Y}(t))}\, dW_j(t), \tag{10}$$

where the $\{W_j(t)\}_{j=1}^{M}$ are independent Brownian motions.

As background for the SDE analysis in section 6, we now state the relevant part of Ito's lemma; see, for example, [23]. For the general Ito SDE system with n components and d independent Brownian motions

$$dY_i(t) = b_i\left(\mathbf{Y}(t)\right) dt + \sum_{j=1}^{d} \sigma_{ij}\left(\mathbf{Y}(t)\right) dW_j(t), \quad 1 \le i \le n, \tag{11}$$

we let

$$a\left(\mathbf{Y}(t)\right) := \boldsymbol{\sigma}\left(\mathbf{Y}(t)\right) \boldsymbol{\sigma}\left(\mathbf{Y}(t)\right)^T \in \mathbb{R}^{n\times n}. \tag{12}$$

Then for any function $f : \mathbb{R}^n \to \mathbb{R}$ that is twice continuously differentiable, Ito's lemma tells us that

$$\begin{aligned} d\,f\left(\mathbf{Y}(t)\right) = &\left(\sum_{i=1}^{n} \frac{\partial f\left(\mathbf{Y}(t)\right)}{\partial x_i} b_i\left(\mathbf{Y}(t)\right) + \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f(\mathbf{Y}(t))}{\partial x_i \partial x_j} a_{ij}\left(\mathbf{Y}(t)\right)\right) dt \\ &+ \text{mart.}, \end{aligned} \tag{13}$$

where "mart." denotes a martingale whose precise form is not relevant to our work. We will use two particular cases of f. When $f\left(\mathbf{Y}\right) = Y_k^2$, (13) becomes

$$d\left(Y_k^2\right) = \left(2\,Y_k\, b_k\left(\mathbf{Y}(t)\right) + a_{kk}\left(\mathbf{Y}(t)\right)\right) dt + \text{mart.} \tag{14}$$

and when $f\left(\mathbf{Y}\right) = Y_k Y_l$, for $k \neq l$, it becomes

$$d\left(Y_k Y_l\right) = \left(Y_l b_k\left(\mathbf{Y}(t)\right) + Y_k b_l\left(\mathbf{Y}(t)\right) + \tfrac{1}{2} a_{kl}\left(\mathbf{Y}(t)\right) + \tfrac{1}{2} a_{lk}\left(\mathbf{Y}(t)\right)\right) dt + \text{mart.} \tag{15}$$

4 Gene Regulation Model

We now consider a model of eukaryotic gene regulation, originally proposed in [25]. This model incorporates two states of promoters: an inactive state, D, not permissive of transcription, and an active state $D^\star$ that is competent for transcription. Transition between the two states of promoter is reversible and the total number of promoters is conserved, i.e. $D + D^\star = D_T$. Transcription takes place from the active state $D^\star$ with the linear rate k_r, resulting in production of messenger RNA (mRNA) molecules that decay with rate γ_r. Proteins P are translated from mRNA molecules with linear rate k_p and they decay with rate γ_p.

This model of gene regulation could be described by the following reactions:

$$D \underset{k_d}{\overset{k_a}{\rightleftharpoons}} D^\star \tag{16}$$

$$D^\star \xrightarrow{k_r} M + D^\star \tag{17}$$

$$M \xrightarrow{\gamma_r} \emptyset \tag{18}$$

$$M \xrightarrow{k_p} P + M \tag{19}$$

$$P \xrightarrow{\gamma_p} \emptyset. \tag{20}$$

A formal and complete specification of the system in terms of stoichiometric vectors and propensity functions can be found in section 6.

It is tempting to reduce the first two reactions (16) that involve two species D and $D^\star$ to just one reaction involving $D^\star$ by exploiting the constraint $D+D^\star = D_T$. We could argue that $D^\star$ is produced with the rate $k_a D_T$ and decays with the rate $(k_a + k_d)D^\star$:

$$\emptyset \xrightarrow{k_a D_T} D^\star$$

$$D^\star \xrightarrow{k_a + k_d} \emptyset.$$

In this formalization, however, we cannot guarantee that once $D^\star = D_T$ no

more production of the active state $D^\star$ will occur. This opens up the possibility of $D^\star > D_T$, which violates the conservation law. Hence, we will work with the full system.

5 Moments for Chemical Master Equation

Gadgil et al. [11] considered generic systems of first-order chemical reactions and derived ODEs that describe the evolution of the first two moments of all species. They split first-order reactions into four categories. The gene transcription model (16)–(20) fits into that framework and involves three of these categories. Reactions $D \rightleftharpoons D^\star$ are of *conversion* type, $D^\star \to M + D^\star$ and $M \to P + M$ are *catalytic*, i.e. the reaction affects one species at a rate that is proportional to some other species, and $M \to \emptyset$ and $P \to \emptyset$ are *degradation* type reactions. The fourth type of reaction in [11], production from a source, is not present in this model.

If we use the symbols $D(t)$, $D^\star(t)$, $M(t)$ and $P(t)$ to denote the number of molecules of each species present at time t, [11, equation (28)] shows that the mean values arising from the CME model satisfy the ODE system

$$\frac{d}{dt}\begin{bmatrix} \mathbb{E}\left[D(t)\right] \\ \mathbb{E}\left[D^\star(t)\right] \\ \mathbb{E}\left[M(t)\right] \\ \mathbb{E}\left[P(t)\right] \end{bmatrix} = \mathcal{K} \cdot \begin{bmatrix} \mathbb{E}\left[D(t)\right] \\ \mathbb{E}\left[D^\star(t)\right] \\ \mathbb{E}\left[M(t)\right] \\ \mathbb{E}\left[P(t)\right] \end{bmatrix}, \quad \text{where } \mathcal{K} = \begin{bmatrix} -k_a & k_d & 0 & 0 \\ k_a & -k_d & 0 & 0 \\ 0 & k_r & -\gamma_r & 0 \\ 0 & 0 & k_p & -\gamma_p \end{bmatrix}. \tag{21}$$

Then introducing a time dependent symmetric matrix $V(t) \in \mathbb{R}^{4\times 4}$ to store the second moments and correlations in the form

$$V(t) := \begin{bmatrix} \mathbb{E}\left[D(t)^2 - D(t)\right] & \mathbb{E}\left[D(t)D^\star(t)\right] & \mathbb{E}\left[D(t)M(t)\right] & \mathbb{E}\left[D(t)P(t)\right] \\ \mathbb{E}\left[D(t)D^\star(t)D(t)\right] & \mathbb{E}\left[D^\star(t)^2 - D^\star(t)\right] & \mathbb{E}\left[M(t)D^\star(t)\right] & \mathbb{E}\left[P(t)D^\star(t)\right] \\ \mathbb{E}\left[M(t)D(t)\right] & \mathbb{E}\left[M(t)D^\star(t)\right] & \mathbb{E}\left[M(t)^2 - M(t)\right] & \mathbb{E}\left[M(t)P(t)\right] \\ \mathbb{E}\left[P(t)D(t)\right] & \mathbb{E}\left[P(t)D^\star(t)\right] & \mathbb{E}\left[M(t)P(t)\right] & \mathbb{E}\left[P(t)^2 - P(t)\right] \end{bmatrix},$$

we may appeal to [11, equation (29)], which says

$$\frac{d}{dt}V(t) = \mathcal{K}V(t) + \left(\mathcal{K}V(t)\right)^T + \Gamma(t) + \Gamma(t)^T, \tag{22}$$

where, in our case, $\Gamma(t) \in \mathbb{R}^{4\times 4}$ has the form

$$\Gamma(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & k_r\mathbb{E}[D^\star(t)] & 0 \\ 0 & 0 & 0 & k_p\mathbb{E}[M(t)] \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Substituting the expressions in (21) for the means, we obtain ODEs for the second moments;

$$\frac{d\,\mathbb{E}[D^2]}{dt} = -2k_a\mathbb{E}[D^2] + 2k_d\mathbb{E}[DD^\star] + k_a\mathbb{E}[D] + k_d\mathbb{E}[D^\star], \tag{23}$$

$$\frac{d\,\mathbb{E}[D^{\star 2}]}{dt} = -2k_d\mathbb{E}[D^{\star 2}] + 2k_a\mathbb{E}[DD^\star] + k_a\mathbb{E}[D] + k_d\mathbb{E}[D^\star], \tag{24}$$

$$\frac{d\,\mathbb{E}[M^2]}{dt} = -2\gamma_r\mathbb{E}[M^2] + 2k_r\mathbb{E}[D^\star M] + k_r\mathbb{E}[D^\star] + \gamma_r\mathbb{E}[M], \tag{25}$$

$$\frac{d\,\mathbb{E}[P^2]}{dt} = -2\gamma_p\mathbb{E}[P^2] + 2k_p\mathbb{E}[MP] + k_p\mathbb{E}[M] + \gamma_p\mathbb{E}[P]. \tag{26}$$

Here, and henceforth, to avoid cluttering the equations we suppress the time dependence, so, for example, $D(t)$ is written simply as D. Similarly, for the correlations we find that

$$\begin{aligned}\frac{d\,\mathbb{E}[DD^\star]}{dt} = &-(k_a + k_d)\mathbb{E}[DD^\star] + k_a\mathbb{E}[D^2] + k_d\mathbb{E}[D^{\star 2}] - k_a\mathbb{E}[D] \\ &- k_d\mathbb{E}[D^\star],\end{aligned} \tag{27}$$

$$\frac{d\,\mathbb{E}[DM]}{dt} = -(k_a + \gamma_r)\mathbb{E}[DM] + k_r\mathbb{E}[DD^\star] + k_d\mathbb{E}[D^\star M], \tag{28}$$

$$\frac{d\,\mathbb{E}[DP]}{dt} = -(k_a + \gamma_p)\mathbb{E}[DP] + k_p\mathbb{E}[DM] + k_d\mathbb{E}[D^\star P], \tag{29}$$

$$\frac{d\,\mathbb{E}[D^\star M]}{dt} = -(k_d + \gamma_r)\mathbb{E}[D^\star M] + k_r\mathbb{E}[D^{\star 2}] + k_a\mathbb{E}[DM], \tag{30}$$

$$\frac{d\,\mathbb{E}[D^\star P]}{dt} = -(k_d + \gamma_p)\mathbb{E}[D^\star P] + k_a\mathbb{E}[DP] + k_p\mathbb{E}[D^\star M], \tag{31}$$

$$\frac{d\,\mathbb{E}[MP]}{dt} = -(\gamma_r + \gamma_p)\mathbb{E}[MP] + k_p\mathbb{E}[M^2] + k_r\mathbb{E}[D^\star P]. \tag{32}$$

6 Moments for Chemical Langevin Equation

The CLE formulation described in subsection 3.2 uses a general state vector $\mathbf{Y}(t)$. For the model (16)–(20), in order to make comparisons easier, we will

re-use the notation from section 5, so that

$$\begin{bmatrix} Y_1(t) \\ Y_2(t) \\ Y_3(t) \\ Y_4(t) \end{bmatrix} =: \begin{bmatrix} D \\ D^\star \\ M \\ P \end{bmatrix}.$$

However, we emphasize that D, $D^\star$, M and P in the CLE are real-valued random variables, whereas those in CME take non-negative integer values. We also emphasize that the time-dependency is not made explicit in this notation.

The stoichiometric vectors for reactions (16)–(20) take the form

$$\boldsymbol{\nu}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

$$\boldsymbol{\nu}_4 = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_5 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\nu}_6 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix},$$

and the propensity functions are $a_1 = k_a D$, $a_2 = k_d D^\star$, $a_3 = k_r D^\star$, $a_4 = \gamma_r M$, $a_5 = k_p M$ and $a_6 = \gamma_p P$. Hence the CLE (10) is an SDE of the form (11) with drift function $\mathbf{b} : \mathbb{R}^4 \to \mathbb{R}^4$ given by

$$\mathbf{b}(\mathbf{Y}(t)) = \begin{bmatrix} -k_a D + k_d D^\star \\ k_a D - k_d D^\star \\ k_r D^\star - \gamma_r M \\ k_p M - \gamma_p P \end{bmatrix} \tag{33}$$

and diffusion function $\boldsymbol{\sigma} : \mathbb{R}^4 \to \mathbb{R}^{4\times 6}$ given by

$$\boldsymbol{\sigma}(\mathbf{Y}(t)) = \begin{bmatrix} -\sqrt{k_a D} & \sqrt{k_d D^\star} & 0 & 0 & 0 & 0 \\ \sqrt{k_a D} & -\sqrt{k_d D^\star} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{k_r D^\star} & -\sqrt{\gamma_r M} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{k_p M} & -\sqrt{\gamma_p P} \end{bmatrix}. \tag{34}$$

So a in (12) satisfies

$$a = \begin{bmatrix} k_a D + k_d D^\star & -k_a D - k_d D^\star & 0 & 0 \\ -k_a D - k_d D^\star & k_a D + k_d D^\star & 0 & 0 \\ 0 & 0 & k_r D^\star + \gamma_r M & 0 \\ 0 & 0 & 0 & k_p M + \gamma_p P \end{bmatrix}. \tag{35}$$

Because the drift coefficient in (33) is linear, taking expectations in the SDE leads to the linear ODE (21) that we obtained for the CME.

Applying Ito's lemma to $f(\mathbf{Y}) = D^2$, using (14) and the expressions in (33) and (35), we find that

$$d\left(D^2\right) = \left(2D\left(-k_a D + k_d D^\star\right) + \left(k_a D + k_d D^\star\right)\right) dt + \text{mart.}$$

so, after taking expectations,

$$\frac{d\,\mathbb{E}[D^2]}{dt} = -2k_a\mathbb{E}[D^2] + 2k_d\mathbb{E}[DD^\star] + k_a\mathbb{E}[D] + k_d\mathbb{E}[D^\star],$$

which matches (23). Similarly, (14) shows that the other second moments satisfy the ODEs (24)–(26). In the same manner, we may apply Ito's lemma to $f(\mathbf{Y}) = DD^\star$, using (15), to find that

$$\begin{aligned} d\left(DD^\star\right) = &\left(D^\star\left(-k_a D + k_d D^\star\right) + D\left(k_a D - k_d D^\star\right)\right. \\ &\left. + \tfrac{1}{2}\left(-k_a D - k_d D^\star - k_a D - k_d D^\star\right)\right) dt + \text{mart.} \end{aligned}$$

So, after taking expectations,

$$\frac{d\,\mathbb{E}[DD^\star]}{dt} = -(k_a + k_d)\mathbb{E}[DD^\star] + k_a\mathbb{E}[D^2] + k_d\mathbb{E}[D^{\star 2}] - k_a\mathbb{E}[D] - k_d\mathbb{E}[D^\star],$$

matching (27). Similarly, the other correlations are found to satisfy (28)–(32).

In summary, the means, variances and correlations for all components satisfy the same ODEs for both the CME and CLE formulations of the model, and hence they are equal for all time.

7 Numerical Experiment for a Bimolecular Case

All reactions in the model (16)–(20) are first-order in the sense of [11]. An important instance where a first-order model is not sufficient arises when proteins produced from the mRNA may combine to form complexes, such as dimers. There is ample experimental evidence to suggest that protein subunits can degrade less rapidly when associated in multimeric complexes, an effect referred to in [4] as "cooperative stability". For dimeric transcription factors, this effect leads to a concentration-dependence in the degradation rate because monomers, which are predominant at low concentrations, will be more rapidly degraded. Thus, cooperative stability can effectively widen the accessible range of protein levels in vivo and a few-fold difference between the degradation rate of monomers and dimers can already enhance the function of these circuits substantially. In [4], the effect of cooperative stability through nonlinear degradation in a simple genetic circuit with feedback was studied without incorporating stochastic effects. On the other hand, SSA simulations were used in [5], but without considering rapid degradation of monomers compared to dimers.

To illustrate a model that incorporates dimerization, we begin with a simplified version of the model in section 4 where there is only a single, active, state of the gene

$$
\begin{aligned}
& \overset{k_r}{\longrightarrow} \\
D^\star &\longrightarrow M + D^\star \\
& \overset{k_p}{\longrightarrow} \\
M &\longrightarrow P + M \\
& \overset{\gamma_r}{\longrightarrow} \\
M &\longrightarrow \emptyset \\
& \overset{\gamma_{p_1}}{\longrightarrow} \\
P &\longrightarrow \emptyset.
\end{aligned}
$$

Then we may allow the protein monomers P to form dimers P_2, which degrade

less rapidly than the monomers ($\gamma_{p_2} < \gamma_{p_1}$):

$$\begin{aligned} & & & k_a & & \\ P &+ P & &\rightleftharpoons & P_2 & \\ & & & k_d & & \\ & & & \gamma_{p_2} & & \\ P_2 & & &\rightarrow & \emptyset. & \end{aligned}$$

Our aim is to test whether the correspondence between first and second moments for the CME and CLE that we proved for the model with unimolecular reactions in section 4 carries through to this case, where a dimerization (and hence second-order) reaction is present. A full Monte Carlo simulation of the CME and CLE would be very expensive (for example, the CLE contains seven independent Brownian motions, so an expected value corresponds to an integral over seven dimensions). Hence, we will focus on a reduced model that contains dimerization. If we assume that the protein arises as production from a source and ignore any possible reversibility of the dimerization, we arrive at the computationally simpler model

$$\emptyset \xrightarrow{k_1} P \tag{36}$$

$$P + P \xrightarrow{k_a} P_2 \tag{37}$$

$$P \xrightarrow{\gamma_p} \emptyset \tag{38}$$

$$P_2 \xrightarrow{\gamma_{p_2}} \emptyset. \tag{39}$$

We emphasize that we are using this model simply to test whether the conclusions of sections 5–6 are close to holding in an example with second order reactions. Writing the state vector as

$$\begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} =: \begin{bmatrix} P \\ P_2 \end{bmatrix},$$

the stoichiometric vectors take the form

$$\boldsymbol{\nu}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \quad \boldsymbol{\nu}_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\nu}_4 = \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

and the propensity functions are $a_1 = k_a$, $a_2 = k_a P(P-1)/2$, $a_3 = \gamma_P P$ and $a_4 = \gamma_{P_2} P_2$.

In this case the CLE takes the form

$$d\begin{bmatrix} P \\ P_2 \end{bmatrix} = \begin{bmatrix} k_1 - k_a P(P-1) - \gamma_P P \\ k_a P(P-1)/2 - \gamma_{P_2} P_2 \end{bmatrix} dt + \begin{bmatrix} \sqrt{k_1}\, dW_1 - \sqrt{k_a P(P-1)}\, dW_2 - \sqrt{\gamma_P P}\, dW_3 \\ \sqrt{k_a P(P-1)/2}\, dW_2 - \sqrt{\gamma_{P_2} P_2}\, dW_4 \end{bmatrix}.$$

In this SDE the first equation shows that P is uncoupled from P_2 (a fact which is also clear from the original formulation (36)–(39)), so we may consider separately the SDE

$$dP = (k_1 - k_a P(P-1) - \gamma_P P)\, dt + \sqrt{k_1}\, dW_1 - \sqrt{k_a P(P-1)}\, dW_2 - \sqrt{\gamma_P P}\, dW_3.$$

Taking expectations leads to an ODE for $\mathbb{E}[P]$ that involves $\mathbb{E}[P^2]$. Similarly, applying Ito's lemma to $f(Y) = P^2$ gives an ODE for $\mathbb{E}[P^2]$ that involves $\mathbb{E}[P^3]$. Because the system is not closed, this does not lead to an analytical formula for the moments, and also hints that our moment matching approach from sections 5 and 6 is unlikely to be successful.

We therefore proceed computationally. Choosing the values $k_1 = 5$, $k_a = 0.01$, $\gamma_P = 0.1$ and $\gamma_{P_2} = 0.01$ with initial conditions $P(0) = 10$ and $P_2(0) = 2$, we consider the time interval $0 \le t \le 20$. For the purposes of illustration, in Figure 3 we show one path for the monomer P and the dimer P_2 from the CME, computed with SSA, and from the CLE, approximated with the Euler–Maruyama method [18,23]. We note that the two computations use different, independent, noise sources (from MATLAB's `rand` and `randn`) and hence there is no reason for the two paths to be close. Then, using Monte Carlo simulations over $K = 10^5$ paths we computed sample mean approximations to $\mathbb{E}[P]$, $\mathbb{E}[P^2]$, $\mathbb{E}[P_2]$ and $\mathbb{E}[P_2^2]$ at time $t = 20$. The results are given in Table 1. Here, we have presented 95% confidence intervals for each sample mean by adding $\pm 1.96\,\mathrm{std}/\sqrt{K}$, where std denotes the sample's standard deviation. All values have been rounded to four significant digits. CLEa denotes the results for Euler–Maruyama using a stepsize $20/500 = 0.04$. The table also shows results for Euler–Maruyama with stepsize 0.004, labeled CLEb, in order to check that numerical discretization errors are not significant.

We see from Table 1 that there is overlap between the computed CME and CLE confidence intervals for both sets of first and second moments. We conclude that, to the typical accuracy obtained from large scale Monte Carlo simulations, the first two moments are indistinguishable.

We conclude with two remarks.

(1) Further computational testing would help to reveal the extent to which moments match for more general models, and in this case some analysis

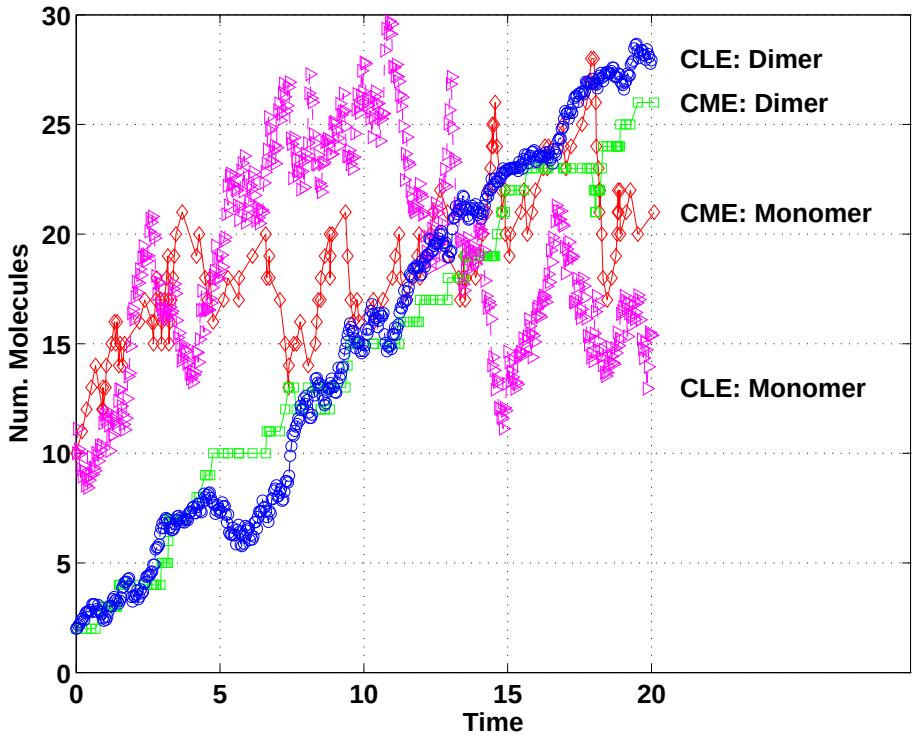

Fig. 3. One path from the Chemical Master Equation (CME) and Chemical Langevin Equation (CLE) for the dimerization model (36)–(39). Diamonds: P from CME; Squares: P_2 from CME; Triangles: P from CLE; Circles: P_2 from CLE.

	$\mathbb{E}[P]$	$\mathbb{E}[P^2]$	$\mathbb{E}[P_2]$	$\mathbb{E}[P_2^2]$
CME	$[17.97, 18.01]$	$[337.3, 339.0]$	$[28.05, 28.10]$	$[806.1, 809.2]$
CLEa	$[17.96, 18.01]$	$[337.0, 338.8]$	$[28.07, 28.13]$	$[807.6, 810.8]$
CLEb	$[17.97, 18.02]$	$[337.4, 339.2]$	$[28.06, 28.11]$	$[807.0, 810.1]$

Table 1
95% confidence intervals for Monte Carlo sample mean approximations to $\mathbb{E}[P]$, $\mathbb{E}[P^2]$, $\mathbb{E}[P_2]$ and $\mathbb{E}[P_2^2]$ at time $t = 20$ in (36)–(39) from the CME and CLE. CLEa uses Euler–Maruyama with stepsize 0.04 and CLEb uses Euler–Maruyama with stepsize 0.004.

might be possible that gives bounds on the discrepancies and indicates parameter regimes where there is a close match.

(2) Generally, there is a need for existence and uniqueness results for the SDEs that can appear in the CLE formulation. The conditions under which the CLE (10) is derived in [15] make it clear that the model is likely to be unrealistic when components $Y_i(t)$ approach zero—this is precisely where the issue of negative arguments inside the square root function rouses itself.

Acknowledgement This work was sponsored by EPSRC grants GR/S62383/01

and GR/S62383/01 (DJH) and by a Synergy grant from the Universities of Strathclyde and Glasgow (RK and DJH). The authors benefited from useful conversations with Dan Gillespie.

References

[1] 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).

[2] 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.

[3] W. Blake, M. Kaern, C. Cantor, and J. Collins, *Noise in eukaryotic gene expression*, Nature, 422 (2003), pp. 633–7.

[4] N. E. Buchler, U. Gerland, and T. Hwa, *Nonlinear protein degradation and the function of genetic circuits*, Proc Natl Acad Sci U S A, 102 (2005), pp. 9559–64.

[5] R. Bundschuh, F. Hayot, and C. Jayaprakash, *The role of dimerization in noise reduction of simple genetic networks*, J. Theor. Biol., 220 (2003), pp. 261–269.

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

[7] J. Chubb, T. Trcek, S. Shenoy, and R. Singer, *Transcriptional pulsing of a developmental gene*, Curr. Biol., 16 (2006), pp. 1018–1025.

[8] E. J. Crampin and S. Schnell, *New approaches to modelling and analysis of biochemical reactions, pathways and networks*, Progress in Biophysics & Molecular Biology, 86 (2004), pp. 1–4.

[9] 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.

[10] H. El 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.

[11] 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.

[12] 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.

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

[14] D. T. Gillespie, *Markov Processes: An Introduction for Physical Scientists*, Academic Press, San Diego, 1991.

[15] D. T. Gillespie, *The chemical Langevin equation*, J. Chem. Phys., 113 (2000), pp. 297–306.

[16] D. T. Gillespie and L. Petzold, *Numerical simulation for biochemical kinetics*, in System Modelling in Cellular Biology: From Concepts to Nuts and Bolts, Z. Szallasi, J. Stelling, and V. Periwal, eds., MIT Press, 2006, pp. 125–147.

[17] I. Golding, J. Paulsson, S. Zawilski, and E. Cox, *Real-time kinetics of gene activity in individual bacteria*, Cell, 123 (2005), pp. 1025–1036.

[18] D. J. Higham, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Review, 43 (2001), pp. 525–546.

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

[20] T. Jahnke and W. Huisinga, *Solving the chemical master equation for monomolecular reaction systems analytically*, Journal of Mathematical Biology, 54 (2007), pp. 1–26.

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

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

[23] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, second ed., 2007.

[24] P. Paszek, *Modeling stochasticity in gene regulation: characterization in the terms of the underlying distribution function*, Bulletin of Mathematical Biology, 69 (2007), pp. 1567–1601.

[25] J. Raser and E. O'Shea, *Control of stochasticity in eukaryotic gene expression*, Science, 304 (2004), pp. 1811–4.

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

[27] P. S. Swain, *Efficient attenuation of stochasticity in gene expression through post-transcriptional control*, J. Mol.Biol., 344 (2004), pp. 965–76.

[28] P. S. Swain, M. B. Elowitz, and E. D. Siggia, *Intrinsic and extrinsic contributions to stochasticity in gene expression*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 12795–800.

[29] T. E. Turner, S. Schnell, and K. Burrage, *Stochastic approaches for modelling in vivo reactions*, Computational Biology and Chemistry, 28 (2004), pp. 165–178.

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