Stochastic Gene Expression: Modeling, Analysis, and Identification*

Mustafa Khammash
Center for Control, Dynamical Systems, and Computations
Department of Mechanical Engineering
Engineering II Bldg. Rm. 2333
University of California at Santa Barbara
Santa Barbara, CA 93106 USA
khammash@engr.ucsb.edu

Brian Munsky
CCS-3 and the Center for NonLinear Studies
Los Alamos National Lab
Los Alamos, NM 87545, USA
brian.munsky@gmail.com

January 6, 2012

Abstract

Gene networks arise due to the interaction of genes through their protein products. Modeling such networks is key to understanding life at the most basic level. One of the emerging challenges to the analysis of genetic networks is that the cellular environment in which these genetic circuits function is abuzz with noise. The main source of this noise is the randomness that characterizes the motion of cellular constituents at the molecular level. Cellular noise not only results in random fluctuations (over time) within individual cells, but it is also a source of phenotypic variability among clonal cellular populations. In some instances fluctuations are suppressed downstream through intricate dynamical networks that act as noise filters. Yet in other important instances, noise induced fluctuations are exploited to the cell's advantage. The richness of stochastic phenomena in biology depends directly upon the interactions of dynamics and noise and upon the mechanisms through which these interactions occur. In this article, we explore the origins and impact of cellular noise, drawing examples from endogenous and synthetic biological networks. We motivate the need for stochastic models and outline the key tools for the modeling and analysis of stochasticity inside living cells. We show that tools from system theory can be effectively utilized for modeling, analysis, and identification of gene networks.

*This article is an expanded version of a conference paper that appeared in the proceedings of IFAC 2009 SYSID

1 Introduction

In living cells, many key events such as gene expression and protein-protein interactions follow from elementary reactions between the cellular consituents at the molecular level (e.g. genes, RNAs, proteins). There is considerable inherent randomness in the order and the timing of these reactions. This randomness can be attributed to the random collisions among cellular constituents whose motion is induced by thermal energy and follows specific statistical distributions. The result is fluctuations in the molecular copy numbers of reaction products both among similar cells and within a single cell over time. These fluctuations (commonly referred to as cellular noise) can propagate downstream–impacting events and processes in accordance to the dynamics of the network interconnection.

Cellular noise has been measured experimentally and classified based on its source ([4, 35]): intrinsic noise refers to noise originating within the boundaries of the process under consideration and is due to the inherent discrete nature of the chemical process of gene expression, whereas extrinsic noise has origins that are more global and affects all processes in the cell under consideration in a similar way (e.g. fluctuations in regulatory protein copy numbers, RNA polymerase numbers, cell-cycle). Noise, both intrinsic and extrinsic, plays a critical role in biological processes. In ([20, 19]) it was proposed that lysis-lysogeny fate decisions for phage λ are determined by a noise driven stochastic switch, implying that the fate of a given cell is determinable only in a probabilistic sense. Another stochastic switch which governs the piliation of E. coli has been modeled in ([22]). Aside from endogenous switches, bistable genetic switches have been constructed and tested ([7, 12]). Depending on their parameters, such switches can be quite susceptible to noise. In ([5]), the first synthetic oscillator was reported. This novel circuit, called the repressilator, consists of three genes, each having a product that represses the next gene, thereby creating a feedback loop of three genes. The role of noise in the operation of the repressilator was recently studied in ([42]). Yet another curious effect of noise can be seen in the fluctuation enhanced sensitivity of intracellular regulation termed 'stochastic focusing' and reported in ([30]). In gene expression, noise induced fluctuations in gene products have been studied in ([37, 38, 33, 14, 36, 1, 31, 29, 17, 2]). Many of these studies look at the propagation of noise in gene networks and the impact (and sometimes limitations) of various types of feedback in suppressing such fluctuations.

In this article, we give an overview of the methods used for modeling and analysis of fluctuations in gene networks. We also demonstrate that these fluctuations can be used in identifying model parameters that may be difficult to measure. The presentation follows that in ([16] and [26]).

1.1 Deterministic vs. Stochastic Modeling

The most common approach for modeling chemical reactions relies on the law of mass-action to derive a set of differential equations that characterize the evolution of reacting species concentrations over time. As an example, consider the reaction $A + B \xrightarrow{k} C$. A deterministic formulation of chemical kinetics would yield the following description $\frac{d[C]}{dt} = k[A] \cdot [B]$ where $[\cdot]$ denotes the concentration, which is considered to be a continuous variable. In contrast, a discrete stochastic formulation of the same reaction describes the *probability* that at a given time, t, the number of molecules of species A and B take certain integer values.

In this way, populations of the species within the network of interest are treated as random variables. In this stochastic description, reactions take place randomly according to certain probabilities determined by several factors including reaction rates and species populations. For example, given certain integer populations of A and B, say N_A and N_B, at time t, the probability that one of the above reactions takes place within the interval $[t, t+dt)$ is proportional to $\frac{N_A \cdot N_B}{\Omega} dt$, where Ω is the volume of the space containing the A and B molecules. In this mesoscopic stochastic formulation of chemical kinetics, molecular species are characterized by their probability density function which quantifies the amount of fluctuations around a certain mean value. As we show below, in the limit of an infinite number of molecules and infinite volume (the thermodynamic limit), fluctuations become negligible and the mesoscopic description converges to the macroscopic description obtained from mass-action kinetics. In typical cellular environments where small volumes and molecule copy numbers are the rule, mesoscopic stochastic descriptions offer a more accurate representations of chemical reactions and their fluctuations. Such fluctuations need to be accounted for as they can generate distinct phenomena that simply cannot be captured by deterministic descriptions. In fact, in certain examples (see e.g. *stochastic focusing* in Fig. 1) the deterministic model fails to even capture the stochastic mean, underscoring the need for stochastic models.

Figure 1: The reaction system shown on the left represents a signaling species S and its response P. I is an intermediate species. When the system is modeled deterministically, the concentration of P fails to capture the stochastic mean of the same species computed from a stochastic model. This example system and the stochastic focusing phenomenon are described in [30].

2 Stochastic Chemical Kinetics

In this section, we provide a more detailed description of the stochastic framework for modeling chemical reactions. In the stochastic formulation of chemical kinetics we shall consider a chemically reacting system of volume Ω containing N molecular species $S_1, ...S_N$ which react via M known reaction channels $R_1....R_M$. We shall make the key assumption that the entire reaction system is well-stirred and is in thermal equilibrium. While this assumption does not always hold in examples of biological networks, spatial models of stochastic chemical kinetics can be formulated. In the well-mixed case that we focus on here, the reaction volume is at a constant temperature T and the molecules move due to the thermal energy. The velocity of a molecule in each of the three spacial directions is independent from the other two and is determined according to a Boltzman distribution:

$$f_{v_x}(v) = f_{v_y}(v) = f_{v_z}(v) = \sqrt{\frac{m}{2\pi k_B T}}\, e^{-\frac{m}{2k_B T}v^2}$$

where m is its mass, v its velocity, and k_B is Boltzman's constant. Let $X(t) = (X_1(t)....X_N(t))^T$ be the state vector, where $X_i(t)$ is a random variable that describes the number of molecules of species S_i in the

system at time t. We consider primitive reactions, which may be either mono-molecular: $S_i \to$ Products, or bi-molecular: $S_i + S_j \to$ Products. More complex reactions can be achieve by introducing intermediate species that interact through a sequence of primitive reactions. In this formulation, each reaction channel R_k defines a transition from some state $\mathbf{X} = \mathbf{x}_i$ to some other state $\mathbf{X} = \mathbf{x}_i + \mathbf{s}_k$, which reflects the change in the state after the reaction has taken place. $\mathbf{s}_k$ is known as the *stoichiometric vector*, and the set of all M reactions give rise to the *stoichiometry matrix* defined as

$$\mathbf{S} = [\mathbf{s}_1 \dots \mathbf{s}_M].$$

Associated with each reaction R_k is a *propensity function*, $w_k(\mathbf{x})$ which captures the rate of the reaction k. Specifically, $w_k(\mathbf{x})dt$ is the probability that, given the system is in state $\mathbf{x}$ at time t, the k^{th} reaction will take place in the time interval $[t, t + dt)$. The propensity function for various reaction types is given in Table 1.

Reaction type	Propensity function
$S_i \to$ Products	$c\mathbf{x}_i$
$S_i + S_j \to$ Products $\quad (i \neq j)$	$c'\mathbf{x}_i\mathbf{x}_j$
$S_i + S_i \to$ Products	$c''\mathbf{x}_i(\mathbf{x}_i - 1)/2$

Table 1: Table showing the propensity function for the various elementary reactions. If we denote by k, k', and k'' the reaction rate constants from deterministic mass-action kinetics for the first, second, and third reaction types shown in the table, it can be shown that $c = k$, $c' = k'/\Omega$, and $c'' = 2k''/\Omega$.

2.1 Sample Path Representation and Connection with Deterministic Models

A sample path representation of the stochastic process $X(t)$ can be given in terms of independent Poisson processes $Y_k(\lambda)$ with parameter λ. In particular, it can be shown [6] that

$$\mathbf{X}(t) = \mathbf{X}(0) + \sum_k \mathbf{s}_k Y_k \left(\int_0^t w_k(\mathbf{X}(s))ds \right).$$

Hence, the Markov process $\mathbf{X}(t)$ can be represented as a random time-change of other Markov processes. When the integral is approximated by a finite sum, the result is an approximate method for generating sample paths which is commonly referred to as tau leaping [10]. The sample path representation shown here is of theoretical interest as well as. Together with the Law of Large numbers, it is used to establish a connection between deterministic and stochastic representations of the same chemical system.

In a deterministic representation based on conventional mass-action kinetics, the solution of the deterministic reaction rate equations arising describes the trajectories of the concentrations of species $S_1, \dots, S_N$.

Let these concentrations be denoted by $\mathbf{\Phi}(t) = [\Phi_1(t), \ldots, \Phi_N(t)]^T$. Accordingly, $\Phi(\cdot)$ satisfies the mass-action ODE:

$$\dot{\mathbf{\Phi}} = \mathbf{S} f(\mathbf{\Phi}(t)), \qquad \mathbf{\Phi}(0) = \mathbf{\Phi}_0.$$

For a meaningful comparison with the stochastic solution, we shall compare the function $\mathbf{\Phi}(t)$ with the volume-normalized stochastic process $\mathbf{X}^\Omega(t) := \frac{\mathbf{X}(t)}{\Omega}$. A natural question is: how does $\mathbf{X}^\Omega(t)$ relate to $\mathbf{\Phi}(t)$? The answer is given by the following fact, which is a consequence of the Law of Large numbers ([6]):

Fact 1 *Let $\mathbf{\Phi}(t)$ be the deterministic solution to the reaction rate equations*

$$\frac{d\mathbf{\Phi}}{dt} = \mathbf{S} f(\mathbf{\Phi}), \ \mathbf{\Phi}(0) = \mathbf{\Phi}_0.$$

Let $\mathbf{X}^\Omega(t)$ be the stochastic *representation of the same chemical systems with $\mathbf{X}^\Omega(0) = \mathbf{\Phi}_0$. Then for every $t \geq 0$:*

$$\lim_{\Omega \to \infty} \sup_{s \leq t} \left| \mathbf{X}^\Omega(s) - \mathbf{\Phi}(s) \right| = 0 \quad \text{almost surely.}$$

To illustrate the convergence of the stochastic system to the deterministic description, we consider a simple one species problem with the following non-linear reaction description:

Reaction	**Stoichiometry**	**Deterministic Description**	**Stochastic Description**
R_1 :	$\emptyset \to S$,	$f_1(x) = 20 + 40\frac{\phi}{40^{10}+\phi^{10}}$,	$w_1(X) = \Omega\left(20 + 40\frac{X/\Omega}{40^{10}+(X/\Omega)^{10}}\right)$
R_2 :	$S \to \emptyset$,	$f_2(x) = \phi$,	$w_2(X) = \Omega\,(X/\Omega)\,.$

From Fig. 2A, which illustrates the production and degradation terms of the reaction rate equation, one can see that the deterministic model has three equilibrium points where these terms are equal. Figs. 2A show the deterministic (smooth) and stochastic (jagged) trajectories of the system from two different initial conditions: $\phi(0) = X(0)/\Omega = 0$ and $\phi(0) = X(0)/\Omega = 100$. and three different volumes $\Omega = \{1, 3, 10\}$. From the plot, it is clear that as the volume increases, the difference between the stochastic and deterministic process shrinks. This is the case for almost every possible initial condition, but with one obvious exception. If the the initial condition were chosen to correspond to the unstable equilibrium, then the deterministic process would remain at equilibrium, but the noise driven stochastic process would not. Of course, this unsteady equilibrium corresponds to a single point of zero measure, thus illustrating the nature of the "almost sure" convergence.

Hence in the thermodynamic limit, the stochastic description converges to the deterministic one. While this result establishes a fundamental connection which ties together two descriptions at two scales, in practice the large volume assumption cannot be justified as the cell volume is fixed, and stochastic descriptions could differ appreciably from their large volume limit.

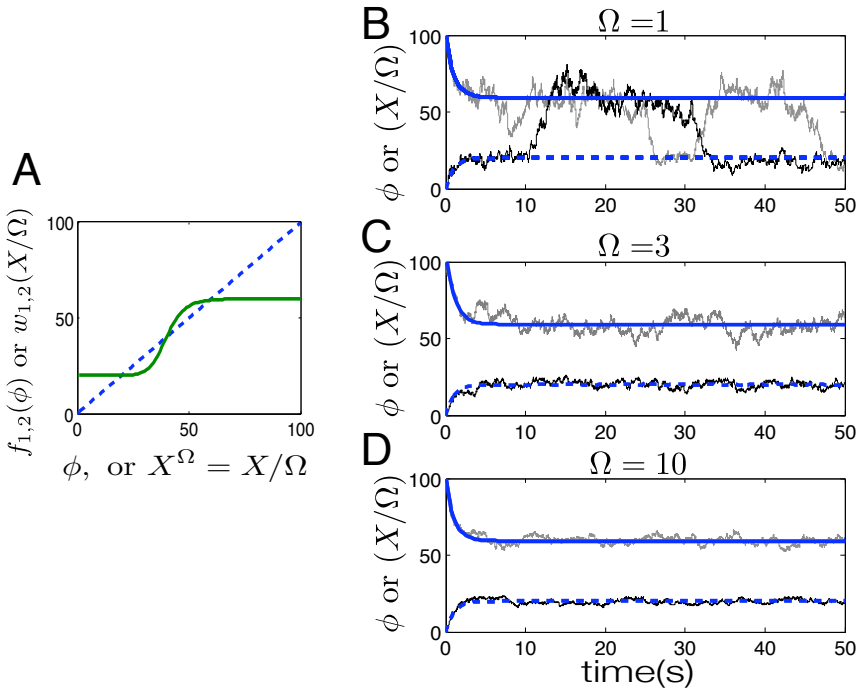

Figure 2: Convergence of the stochastic and deterministic descriptions with volume scaling. (A) Reaction rates for the production (solid) and degradation (dashed) events. (B-D) Trajectories of the deterministic (smooth) and stochastic representations (jagged) assuming equivalent initial conditions for different volumes: (B) $\Omega = 1$, (C) $\Omega = 3$, (D) $\Omega = 10$.

2.2 The Forward Kolmogorov Equation

The Chemical Master Equation (CME), or the Forward Kolmogorov equation, describes the time-evolution of the probability that the chemical reaction system is in any given state, say $\mathbf{X}(t) = \mathbf{x}$. The CME can be derived from the Markov property of chemical reactions. Let $P(\mathbf{x}, t)$, denote the probability that the system is in state $\mathbf{x}$ at time t. We can express $P(\mathbf{x}, t + dt)$ as follows:

$$P(\mathbf{x}, t + dt) \quad = \quad P(\mathbf{x}, t)(1 - \sum_k w_k(\mathbf{x})dt) + \sum_k P(\mathbf{x} - \mathbf{s}_k, t)w_k(\mathbf{x} - \mathbf{s}_k)dt + \mathcal{O}(dt^2).$$

The first term on the right hand side is the probability that the system is already in state $\mathbf{x}$ at time t and no reactions occur to change that in the next dt. In the second term on the right hand side, the kth term in the summation is the probability that the system at time t is an R_k reaction away from being at state $\mathbf{x}$, and that an R_k reaction takes place in the next dt.

Moving $P(\mathbf{x}, t)$ to the left hand side, dividing by dt, and taking the limit as dt goes to zero we get the Chemical Master Equation (CME):

$$\frac{dP(\mathbf{x}, t)}{dt} = \sum_{k=1}^{M} [w_k(\mathbf{x} - \mathbf{s}_k)P(\mathbf{x} - \mathbf{s}_k, t) - w_k(\mathbf{x})P(\mathbf{x}, t)]$$

3 Stochastic Analysis Tools

Stochastic analysis tools may be broadly divided into four categories. The first consists of Kinetic Monte Carlo methods which compute sample paths whose statistics are used to extract information about the system. The second class of methods consists of approximations of the stochastic process $\mathbf{X}(t)$ by solutions of certain stochastic differential equations. The third type of methods seek to compute the trajectories of various moments of $\mathbf{X}(t)$, while the fourth type is concerned with computing the evolution of probability densities of the stochastic process $\mathbf{X}(t)$.

3.1 Kinetic Monte Carlo Simulations

Because the CME is often infinite dimensional, the majority of analyses at the mesoscopic scale have been conducted using Kinetic Monte Carlo algorithms. The most widely used of these algorithms is Gillespie's Stochastic Simulation Algorithm (SSA) [9] and its variants. These are described next.

3.1.1 The Gillespie Algorithm

Each step of Gillespie's SSA begins at a time t and at a state $\mathbf{X}(t) = \mathbf{x}$ and is comprised of three substeps: (*i*) generate the time until the next reaction; (*ii*) determine which reaction occurs at that time; and (*iii*) update the time and state to reflect the previous two choices. The SSA approach is exact in the sense that it results in a random variable with a probability distribution exactly equal to the solution of the corresponding CME. However, each run of the SSA provides only a single trajectory. Numerous trajectories are generated which are then used to compute statistics of interest.

We now describe these steps in more detail. To each of the reactions $\{R_1, \ldots, R_M\}$ we associate a random variable $\mathcal{T}_i$ which describes the time for the next firing of reaction R_i. A key fact is that $\mathcal{T}_i$ is exponentially distributed with parameter w_i. From these, we can define two additional random variables, one continuous and the other discrete:

$$\mathcal{T} = \min_i\{\mathcal{T}_i\} \qquad \text{(Time to the next reaction)}$$
$$\mathcal{R} = \arg\min_i\{\mathcal{T}_i\} \qquad \text{(Index of the next reaction)}$$

It can be shown that: (a) $\mathcal{T}$ is exponentially distributed with parameter: $\sum_i w_i$; and (b) $\mathcal{R}$ has the discrete distribution: $P(\mathcal{R} = k) = \frac{w_k}{\sum_i w_i}$. With this in mind, we are ready to give the steps in Gillespie's Stochastic Simulation Algorithm.

Gillespie's SSA:

- **Step 0** Initialize time t and state population $\mathbf{x}$.
- **Step 1** Draw a sample τ from the distribution of $\mathcal{T}$. (See Figure 3).
- **Step 2** Draw a sample μ from the distribution of $\mathcal{R}$. (See Figure 3).
- **Step 3** Update time: $t \leftarrow t + \tau$. Update the state: $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{s}_\mu$.

Figure 3: The figure shows the cumulative distribution of the two random variables $\mathcal{T}$ and $\mathcal{R}$. A sample of $\mathcal{T}$ is drawn by first drawing a uniformly distributed random number r_1 and then finding its inverse image under F, the cumulative distribution of $\mathcal{T}$. A similar procedure can be used to draw a sample from the distribution of $\mathcal{R}$.

3.2 Stochastic Differential Equation Approximations

There are several SDE approximations of the stochastic process $\mathbf{X}(t)$. One of these is the so-called *Chemical Langevin Equation*, also called the *Diffusion Approximation* ([18, 8]). We will not discuss this here, but we will instead examine another SDE approximation that relates to SDEs that arise naturally in systems and control settings.

Another approximation that leads to a stochastic differential equation is the so called van Kampen's approximation or Linear Noise Approximation (LNA) (see [40, 3, 39]). It is essentially an approximation to the process $\mathbf{X}(t)$ that takes advantage of the fact that in the large volume limit ($\Omega \to \infty$), the process $X^\Omega(t) := \mathbf{X}(t)/\Omega$ converges to the solution $\mathbf{\Phi}(t)$ of the deterministic reaction rate equation: $\dot{\mathbf{\Phi}}(t) = f(\mathbf{\Phi})$. Defining a scaled "error" process $\mathbf{V}^\Omega(t) := \sqrt{\Omega}\big(\mathbf{X}^\Omega(t) - \mathbf{\Phi}(t)\big)$ and using the Central limit theorem, it can be shown that $V^\Omega(t)$ converges in distribution to the solution $V(t)$ to the following linear stochastic differential equation:

$$d\mathbf{V}(t) = \mathbf{J}_f(\mathbf{\Phi})\mathbf{V}(t)dt + \sum_{k=1}^{M} \mathbf{s}_k \sqrt{w_k(\mathbf{\Phi})} d\mathbf{B}_k(t),$$

where $\mathbf{J}_f$ denotes the Jacobian of $f(\cdot)$ ([6]). Hence, the LNA results in a state $\mathbf{X}(t) \approx \Omega\mathbf{\Phi}(t) + \sqrt{\Omega}\mathbf{V}(t)$, which can be viewed as the sum of a deterministic term given by the solution to the deterministic reaction rate equation, and a zero mean stochastic term given by the solution to a linear SDE. While the LNA is

reasonable for systems with sufficiently large numbers of molecules (and volume), examples show that it can yield poor results when this assumption is violated, e.g. when the system of interest contains species with very small molecular counts, or where the reaction propensity functions are strongly nonlinear over the dominant support region of the probability density function.

3.3 Statistical Moments

When studying stochastic fluctuations that arise in gene networks, one is often interested in computing moments and variances of noisy expression signals. The evolution of moment dynamics can be described using the Chemical Master Equation. To compute the first moment $\mathbb{E}[X_i]$, we multiply the CME by x_i and then sum of all $(x_1, \ldots, x_N) \in \mathbb{N}^N$ to get

$$\frac{d\mathbb{E}[X_i]}{dt} = \sum_{k=1}^{M} s_{ik}\mathbb{E}[w_k(X)]$$

Similarly, to get the second moments $\mathbb{E}[X_iX_j]$, we multiply the CME by x_ix_j and sum over all $(x_1, \ldots, x_N) \in \mathbb{N}^N$, which gives

$$\frac{d\mathbb{E}[X_iX_j]}{dt} = \sum_{k=1}^{M} s_{ik}\mathbb{E}[X_jw_k(X)] + \mathbb{E}[X_iw_k(X)]s_{jk} + s_{ik}s_{jk}\mathbb{E}[w_k(X)]$$

These last two equations can be expressed more compactly in matrix form. Defining $\mathbf{w}(x) = [w_1(x), \ldots, w_M(x)]^T$, the moment dynamics become:

$$\begin{aligned}
\frac{d\mathbb{E}[\mathbf{X}]}{dt} &= \mathbf{S}\mathbb{E}[\mathbf{w}(\mathbf{X})] \\
\frac{d\mathbb{E}[\mathbf{X}\mathbf{X}^T]}{dt} &= \mathbf{S}\mathbb{E}[\mathbf{w}(\mathbf{X})\mathbf{X}^T] + \mathbb{E}[\mathbf{w}(\mathbf{X})\mathbf{X}^T]^T\mathbf{S}^T + \mathbf{S}\{diag\mathbb{E}[\mathbf{w}(\mathbf{X})]\}\mathbf{S}^T
\end{aligned}$$

In general, this set of equations cannot be solved explicitly. This is because the moment equations will not always be closed: depending on the form of the propensity vector $w(\cdot)$, the dynamics of the first moment $\mathbb{E}(\mathbf{X})$ may depend on the second moments $\mathbb{E}(\mathbf{X}\mathbf{X}^T)$, the second moment dynamics may in turn depend on the third moments, etc. resulting in an infinite system of ODE's. However, when the propensity function is affine, i.e. $\mathbf{w}(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{w}_0$, where $\mathbf{W}$ is $N \times N$ and $\mathbf{w}_0$ is $N \times 1$, then $\mathbb{E}[\mathbf{w}(\mathbf{X})] = \mathbf{W}\mathbb{E}[\mathbf{X}] + \mathbf{w}_0$, and $\mathbb{E}[\mathbf{w}(\mathbf{X})\mathbf{X}^T] = \mathbf{W}\mathbb{E}[\mathbf{X}\mathbf{X}^T] + \mathbf{w}_0\mathbb{E}[\mathbf{X}^T]$. This gives us the following moment equations:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[\mathbf{X}] &= \mathbf{S}\mathbf{W}\mathbb{E}[\mathbf{X}] + \mathbf{S}\mathbf{w}_0 \\
\frac{d}{dt}\mathbb{E}[\mathbf{X}\mathbf{X}^T] &= \mathbf{S}\mathbf{W}\mathbb{E}[\mathbf{X}\mathbf{X}^T] + \mathbb{E}[\mathbf{X}\mathbf{X}^T]\mathbf{W}^T\mathbf{S}^T + \mathbf{S}\ diag(\mathbf{W}\mathbb{E}[\mathbf{X}] + \mathbf{w}_0)\mathbf{S}^T + \mathbf{S}\mathbf{w}_0\mathbb{E}[\mathbf{X}^T] + \mathbb{E}[\mathbf{X}]\mathbf{w}_0^T\mathbf{S}^T
\end{aligned}$$

Clearly, this is a closed system of linear ODEs that can be solved easily for the first and second moments.

Defining the covariance matrix $\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}(\mathbf{X})]^T]$, we can also compute covariance equations:

$$\frac{d}{dt}\mathbf{\Sigma} = \mathbf{S}\mathbf{W}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{W}^T\mathbf{S}^T + \mathbf{S}\ diag(\mathbf{W}\mathbb{E}[\mathbf{X}] + \mathbf{w}_0)\mathbf{S}^T$$

The steady-state moments and covariances can be obtained by solving linear algebraic equations. Let $\bar{\mathbf{X}} = \lim_{t\to\infty} \mathbb{E}[\mathbf{X}(t)]$ and $\bar{\mathbf{\Sigma}} = \lim_{t\to\infty} \mathbf{\Sigma}(t)$. Then

$$\mathbf{S}\mathbf{W}\bar{\mathbf{X}} = -\mathbf{S}\mathbf{w}_0$$

$$\mathbf{S}\mathbf{W}\bar{\mathbf{\Sigma}} + \mathbf{\Sigma}\bar{\mathbf{W}}^T\mathbf{S}^T + \mathbf{S}\ diag(\mathbf{W}\bar{\mathbf{X}} + \mathbf{w}_0)\mathbf{S}^T = 0$$

The latter is an algebraic Lyapunov equation can be solved efficiently.

3.3.1 Moment Closures.

An important property of the Markov processes that describe chemical reactions is that when one constructs a vector μ with all the first and second-order statistical uncentered moments of the process' state $\mathbf{X}$, this vector evolves according to a *linear* equation of the form

$$\dot{\mu} = \mathbf{A}\mu + \mathbf{B}\bar{\mu}. \tag{1}$$

Unfortunately, as pointed out earlier, (1) is not in general a closed system because the vector $\bar{\mu}$ may contain moments of order larger than two, whose evolution is not provided by (1). In fact, this will always be the case when bi-molecular reactions are involved. A technique that can be used to overcome this difficulty consists of approximating the *open linear* system (1) by a *closed nonlinear* system

$$\dot{\nu} = \mathbf{A}\nu + \mathbf{B}\varphi(\nu), \tag{2}$$

where ν is an approximation to the solution μ to (1) and $\varphi(\cdot)$ is a *moment closure function* that attempts to approximate the moments in $\bar{\mu}$ based on the values of the moments in μ. The construction of $\varphi(\cdot)$ often relies on postulating a given type for the distribution of X and then expressing the higher-order moments in $\bar{\mu}$ by a nonlinear function $\varphi(\mu)$ of the first and second-order moments in μ. Authors construct moment closure functions $\varphi(\cdot)$ based on different assumed distributions for $\mathbf{X}$, which include normal ([41, 28, 11]), lognormal ([15, 34]), Poisson, and binomial ([27]). Here we discuss only the normal and lognormal moment closure method.,

1. *Normal Distribution.* Assuming that the populations of each species follow a multi-variate normal distribution leads to the equation:

$$\mathbb{E}([X_i - \mathbb{E}(X_i)][X_j - \mathbb{E}(X_j)][X_k - \mathbb{E}(X_k)]) = 0$$

from which an expression for the third order moment $\mathbb{E}[X_iX_jX_k]$ in terms of lower order moments can be obtained. When substituted in the moment equations (1), a closed-system results. This is referred to as the Mass-Fluctuation Kinetics in ([11]). So long as the reaction rates are at most second order, only the expressions for the third moments will be necessary–all of which can be determined as above. For third or higher order propensity functions, the resulting higher order moments can be also be easily expressed in terms of the first two using moment using generating functions as described in the example section below.

2. *Lognormal Distribution.* Based on a lognormal distribution for $\mathbf{X}$, one obtains the following equation:

$$\mathbb{E}[X_iX_jX_k] = \frac{\mathbb{E}[X_iX_j]\mathbb{E}[X_jX_k]\mathbb{E}[X_iX_k]}{\mathbb{E}[X_i]\mathbb{E}[X_j]\mathbb{E}[X_k]}.$$

As before, this leads to a closed-system when substituted in the moment Eqn. (1), provided that the reactions in the system are at most bimolecular. In ([13]) it was shown that this moment closure results without any *a priori* assumptions on the shape of the distribution for $\mathbf{X}$ by matching all (or a large number of) the time derivatives of the exact solution for Eqn. (1) with the corresponding time derivatives of the approximate solution for Eqn. (2), for a given set of initial conditions. However, for systems with third or higher order terms in the reaction rates, it is more difficult to find expressions for the higher moments necessary to close the system.

When the population standard deviations are not much smaller than the means, choosing $\varphi(\cdot)$ based on a normal distribution assumption often leads to less accurate approximations. Furthermore, normal distributions of $\mathbf{X}$ allows for negative values of $\mathbf{X}$, which clearly does not reflect the positive nature of the populations represented by $\mathbf{X}(t)$. In these cases, a lognormal or other positive distribution closure may be preferred, but at the cost of more complicated closure expressions for the higher order moments.

3.4 Density Computations

Another approach to analyze models described by the CME aims to compute the probability density functions for the random variable $\mathbf{X}$. This is achieved by approximate solutions of the CME, using a new analytical approach called the Finite State Projection (FSP) ([23, 32, 24, 21]). The FSP approach relies on a projection that preserves an important subset of the state space (e.g. that supporting the bulk of the probability distribution), while projecting the remaining large or infinite states onto a single 'absorbing' state. See Figure 4.

Probabilities for the resulting finite state Markov chain can be computed exactly, and can be shown to give a lower bound for the corresponding probability for the original full system. The FSP algorithm provides a means of systematically choosing a projection of the CME, which satisfies any prespecified accuracy requirement. The basic idea of the FSP is as follows. In matrix form, the CME may be written as $\dot{\mathbf{P}}(t) = \mathbf{A}\mathbf{P}(t)$, where $\mathbf{P}(t)$ is the (infinite) vector of probabilities corresponding to each possible state in the configuration space. The generator matrix $\mathbf{A}$ embodies the propensity functions for transitions from one configuration to another and is defined by the reactions and the enumeration of the configuration

Figure 4: The Finite State Projection. Left panel shows the state space for a system with two species. Arrows indicate possible transitions within states. The corresponding process is a continuous-time discrete state Markov process whose state space is typically very large or infinite. Right panel shows the projected system for a specific projection region (gray box). The projected system is obtained as follows: Transitions within the projection region are kept unchanged. Transitions that emanate from states within the region and end at states outside (in the original system) are routed to a single absorbing state in the projected system. Transitions into the projection region are deleted. As a result, the projected system is a finite state Markov process, and the probability of each state can be computed exactly.

space. A projection can now be made to achieve an arbitrarily accurate approximation as outlined next: Given an index set of the form $J = \{j_1, j_2, j_3, \ldots\}$ and a vector $\mathbf{v}$, let $\mathbf{v}_J$ denote the subvector of $\mathbf{v}$ chosen according to J, and for any matrix $\mathbf{A}$, let $\mathbf{A}_J$ denote the submatrix of $\mathbf{A}$ whose rows and columns have been chosen according to J. With this notation, we can restate the result from [23, 21]: *Consider any distribution which evolves according to the linear ODE* $\dot{\mathbf{P}}(t) = \mathbf{A}\mathbf{P}(t)$. *Let* $\mathbf{A}_J$ *be a principle sub-matrix of* $\mathbf{A}$ *and* $\mathbf{P}_J$ *be a sub-vector of* $\mathbf{P}$, *both corresponding to the indexes in* J. *If for a given* $\varepsilon > 0$ *and* $t_f \geq 0$ *we have that* $\mathbf{1}^T \exp(A_J t_f)\mathbf{P}_J(0) \geq 1 - \varepsilon$, *then*

$$\| \exp(\mathbf{A}_J t_f)\mathbf{P}_J(0) - \mathbf{P}_J(t_f)\|_1 \leq \varepsilon,$$

which provides a bound on the error between the exact solution $\mathbf{P}_J$ to the (infinite) CME and the matrix exponential of the (finite) reduced system with generator $\mathbf{A}_J$. This result is the basis for an algorithm to compute the probability density function with guaranteed accuracy. The FSP approach and various improvements on the main algorithm can be found in [24, 21].

4 Parameter Identification

Microscopy techniques and Fluorescence Activated Cell Sorting (FACS) technology enable single cell measurement of cellular species to be carried out for large numbers of cells. This raises the prospect of using statistical quantities such as moments and variances, measured at different instants in time, to identify model parameters. Here we demonstrate these ideas through a simple description of gene transcription and translation. Let x denote the population of mRNA molecules, and let y denote the population of proteins in a cell. The system population is assumed to change only through four reactions:

$$\emptyset \rightarrow mRNA$$

$$mRNA \rightarrow \emptyset$$

$$mRNA \rightarrow mRNA + protein$$

$$protein \rightarrow \emptyset$$

for which the propensity functions, $w_i(x, y)$, are

$$w_1(x, y) = k_1 + k_{21}y; \quad w_2(x, y) = \gamma_1 x;$$

$$w_3(x, y) = k_2 x; \quad w_2(x, y) = \gamma_2 y.$$

Here, the terms k_i and γ_i are production and degradation rates, respectively, and k_{21} corresponds to a feedback effect that the protein is assumed to have on the transcription process. In positive feedback, $k_{21} > 0$, the protein increases transcription; in negative feedback, $k_{21} < 0$, the protein inhibits transcription.

The various components of the first two moments, $\mathbf{v}(t) := \left[\begin{array}{ccccc} \mathbb{E}\{x\} & \mathbb{E}\{x^2\} & \mathbb{E}\{y\} & \mathbb{E}\{y^2\} & \mathbb{E}\{xy\} \end{array} \right]^T$, evolve according to the linear time invariant system:

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} \mathbb{E}\{x\} \\ \mathbb{E}\{x^2\} \\ \mathbb{E}\{y\} \\ \mathbb{E}\{y^2\} \\ \mathbb{E}\{xy\} \end{bmatrix} &= \begin{bmatrix} -\gamma_1 & 0 & k_{21} & 0 & 0 \\ \gamma_1 + 2k_1 & -2\gamma_1 & k_{21} & 0 & 2k_{21} \\ k_2 & 0 & -\gamma_2 & 0 & 0 \\ k_2 & 0 & \gamma_2 & -2\gamma_2 & 2k_2 \\ 0 & k_2 & k_1 & k_{21} & -\gamma_1 - \gamma_2 \end{bmatrix} \begin{bmatrix} \mathbb{E}\{x\} \\ \mathbb{E}\{x^2\} \\ \mathbb{E}\{y\} \\ \mathbb{E}\{y^2\} \\ \mathbb{E}\{xy\} \end{bmatrix} + \begin{bmatrix} k_1 \\ k_1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\ &= \mathbf{A}\mathbf{v} + \mathbf{b} \end{aligned} \tag{3}$$

Now that we have expressions for the dynamics of the first two moments, they can be used to identify the various parameters: $[k_1, \gamma_1, k_2, \gamma_2, k_{21}]$ from properly chosen data sets. We will next show how this can be done for transcription parameters k_1 and γ_1. For a discussion on identification of the full set, we refer the reader to ([26, 21, 25]).

4.1 Identifying Transcription Parameters

We begin by considering a simpler birth-death process of mRNA transcripts, whose populations are denoted by x. The moment equation for this system is:

$$\frac{d}{dt} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -\gamma & 0 \\ \gamma + 2k & -2\gamma \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + \begin{bmatrix} k \\ k \end{bmatrix},$$

where we have dropped the subscripts on k_1 and γ_1. By applying the nonlinear transformation:

$$\begin{bmatrix} \mu \\ \sigma^2 - \mu \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 - v_1^2 - v_1 \end{bmatrix},$$

where μ and σ^2 refer to the mean and variance of x, respectively, we arrive at the transformed set of equations:

$$\begin{aligned}\frac{d}{dt}\begin{bmatrix} \mu \\ \sigma^2 - \mu \end{bmatrix} &= \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 - 2v_1\dot{v}_1 - \dot{v}_1 \end{bmatrix} \\ &= \begin{bmatrix} -\gamma_1 v_1 + k \\ (\gamma_1 + 2k)v_1 - 2\gamma v_2 + k - (2v_1 + 1)(-\gamma v_1 + k) \end{bmatrix} \\ &= \begin{bmatrix} -\gamma & 0 \\ 0 & -2\gamma \end{bmatrix}\begin{bmatrix} \mu \\ \sigma^2 - \mu \end{bmatrix} + \begin{bmatrix} k \\ 0 \end{bmatrix}. \end{aligned} \tag{4}$$

Suppose that μ and σ^2 are known at two instances in time, t_0 and $t_1 = t_0 + \tau$, and denote their values at time t_i as μ_i and σ_i^2, respectively. The relationship between (μ_0, σ_0^2) and (μ_1, σ_1^2) is governed by the solution of (4), which can be written:

$$\begin{bmatrix} \mu_1 \\ \sigma_1^2 - \mu_1 \end{bmatrix} = \begin{bmatrix} \exp(-\gamma\tau)\mu_0 \\ \exp(-2\gamma\tau)(\sigma_0^2 - \mu_0) \end{bmatrix} + \begin{bmatrix} \frac{k}{\gamma}(1 - \exp(-\gamma\tau)) \\ 0 \end{bmatrix} \tag{5}$$

In this expression there are 2 unknown parameters, γ and k, that we wish to identify from the data $\{\mu_0, \sigma_0^2, \mu_1, \sigma_1^2\}$. If $\mu_0 = \sigma_0^2$, the second equation is trivial, and we are left with only one equation whose solution could be any pair:

$$\left(\gamma,\ k = \gamma\frac{\mu_1 - \exp(-\gamma\tau)\mu_0}{1 - \exp(-\gamma\tau)}\right).$$

If for the first measurement $\mu_0 \neq \sigma_0^2$ and for the second measurement $\mu_1 \neq \sigma_1^2$, then we can solve for:

$$\begin{aligned}\gamma &= -\frac{1}{2t}\log\left(\frac{\sigma_1^2 - \mu_1}{\sigma_0^2 - \mu_0}\right) \\ k &= \gamma\frac{\mu_1 - \exp(-\gamma t)\mu_0}{1 - \exp(-\gamma\tau)}.\end{aligned}$$

Note that if μ_1 and σ_1^2 are very close, the sensitivity of γ to small errors in this difference becomes very large. From (5), one can see that as τ becomes very large, $(\sigma_1^2 - \mu_1)$ approaches zero, and *steady state measurements do not suffice to uniquely identify both parameters.*

5 Examples

To illustrate the above methods, we consider the synthetic self regulated genetic system as illustrated in Fig. 5. The *lac* operon controls the production of the LacI protein, which in turn tetramerizes and represses its own production. The *lac* operon is assumed to be present in only a single copy within each cell and is assumed to have two possible state: g_{ON} and g_{OFF}, which are characterized by whether or not a LacI

tertramer, LacI_4 is bound to the operon. In all, the model is described with seven reactions:

$$\begin{array}{ccc} \text{Reaction\#} & \text{Reaction Description} & \text{Propensity Function} \\ \hline R1: & 4\text{LacI} \rightarrow \text{LacI}_4 & w_1 = k_1 \binom{[\text{LacI}]}{4} \\ R2: & \text{LacI}_4 \rightarrow 4\text{LacI} & w_2 = k_2[\text{LacI}_4] \\ R3: & g_{ON} + \text{LacI}_4 \rightarrow g_{OFF} & w_3 = k_3[g_{ON}][\text{LacI}_4] \\ R4: & g_{OFF} \rightarrow \text{LacI}_4 + g_{ON} & w_4 = k_4[g_{OFF}] \\ R5: & g_{ON} \rightarrow g_{ON} + \text{LacI} & w_5 = k_5[g_{ON}] \\ R6: & \text{LacI} \rightarrow \phi & w_6 = k_6[\text{LacI}] \\ R7: & \text{LacI}_4 \rightarrow \phi & w_7 = k_7[\text{LacI}_4] \end{array} \tag{6}$$

The first of these reactions corresponds to the combination of four individual monomers to form a tetramer– the rate of this reaction depends upon the total number of possible combinations of four different molecules, which is given by the binomial

$$\binom{[\text{LacI}]}{4} = [\text{LacI}] \cdot ([\text{LacI}] - 1) \cdot ([\text{LacI}] - 2) \cdot ([\text{LacI}] - 3)/24,$$

and the second reaction corresponds to the reverse of the tetramerization event. The next two reactions characterize the ON-to-OFF and OFF-to-ON switches that occur when a tetramer binds to or unbinds from the operon, respectively. When the gene is in the ON state, then the fifth reaction can occur and LacI monomers are created with an exponentially distributed waiting times. Finally reactions R6 and R7 correspond to the usual linear decay of the monomers and tetramers, respectively.

For the analysis of this process, we first define the stoichiometry and reaction rate vectors for the process as:

$$\mathbf{S} = \begin{bmatrix} -4 & 4 & 0 & 0 & 1 & -1 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}, \text{ and} \tag{7}$$

$$\mathbf{w}(\mathbf{x}) = \begin{bmatrix} k_1 \binom{x_1}{4} \\ k_2 x_2 \\ w_3 = k_3 x_3 x_2 \\ w_4 = k_4 x_4 \\ w_5 = k_5 x_3 \\ w_6 = k_6 x_1 \\ w_7 = k_7 x_2 \end{bmatrix}. \tag{8}$$

In what follows, we will take many different approaches to analyzing this system. In order to compare each method, we make the assumption that the volume is unity $\Omega = 1$, such that we can avoid parameter

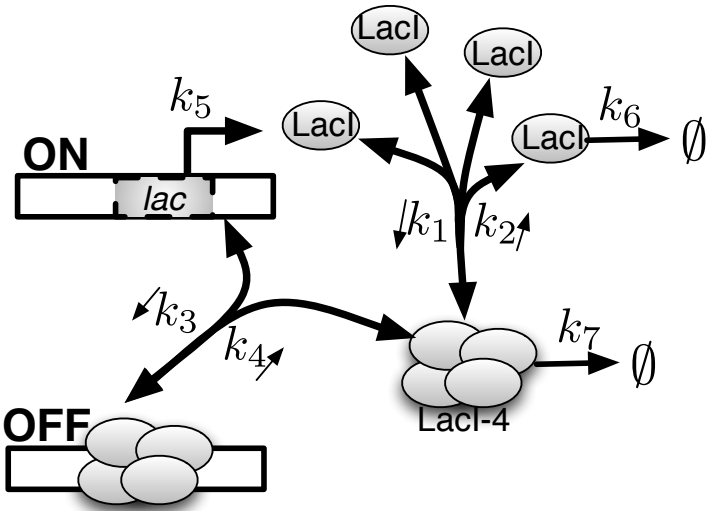

Figure 5: Schematic representation of a synthetic self regulated genetic network. In the model, four LacI monomers (represented as ovals) can bind reversibly to form tetramers (represented as clusters of four ovals). The *lac* operon has two states: OFF corresponding to when LacI tetramers are bound to the gene and blocking the transcription start site, and ON when LacI tetramers are not bound to the gene. Both LacI monomers and tetramers can degrade. See also reactions listed in Eq. 6.

scaling issues when moving between reaction rate equations and the stochastic description. We consider the following parameter set for the reaction rates:

$$\begin{array}{lll} k_1 = 1/30 \text{ N}^{-4}\text{s}^{-1} & k_2 = 0.002 \text{ N}^{-1}\text{s}^{-1} & k_3 = 0.01 \text{ N}^{-2}\text{s}^{-1} \\ k_4 = 0.2 \text{ N}^{-1}\text{s}^{-1} & k_5 = 20 \text{ N}^{-1}\text{s}^{-1} & k_6 = 0.1 \text{ N}^{-1}\text{s}^{-1} \\ k_7 = 0.1 \text{ N}^{-1}\text{s}^{-1}, & & \end{array}$$

and we assume that the process begins with the gene in the active state and no LacI is present in the system:

$$\mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \\ x_4(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

5.1 Deterministic (Reaction Rate) Analysis

As a first analysis, let us consider the deterministic reaction rate equations that are described by these four interacting chemical species and their seven reactions. For this case, one can write the reaction rate

equations as:

$$\dot{\mathbf{x}}(t) = \mathbf{S}\mathbf{w}(\mathbf{x}(t))$$

or in the usual notation of ordinary differential equations:

$$\begin{aligned}
\dot{x}_1 &= -(4/24)k_1x_1(x_1-1)(x_1-2)(x_1-3) + 4k_2x_2 + k_5x_3 - k_6x_1,\\
\dot{x}_2 &= (4/24)k_1x_1(x_1-1)(x_1-2)(x_1-3) - k_2x_2 - k_3x_2x_3 - k_7x_2,\\
\dot{x}_3 &= -k_3x_2x_3 + k_4x_4,\\
\dot{x}_4 &= k_3x_2x_3 - k_4x_4.
\end{aligned}$$

We note that the first reaction only makes sense when the $x_1 \geq 4$ corresponding to when there are at least four molecules of the monomer present and able to combine. In the case where there are fewer than four molecules, we must use a different set of equations:

$$\begin{aligned}
\dot{x}_1 &= 4k_2x_2 + k_5x_3 - k_6x_1,\\
\dot{x}_2 &= -k_2x_2 - k_3x_2x_3 - k_7x_2,\\
\dot{x}_3 &= -k_3x_2x_3 + k_4x_4,\\
\dot{x}_4 &= k_3x_2x_3 - k_4x_4.
\end{aligned}$$

These equations have been integrated over time and the responses of the dynamical process are shown in the solid gray lines of Fig. 7. We note that were one to use the linear noise approximation, the computed mean value for the process would be exactly the same as the solutions shown with the solid gray line.

5.2 Stochastic Simulations

The reactions listed above can also be simulated using Gillespie's stochastic simulation algorithm (SSA-[9]. Two such simulations shown in Fig. 6 illustrate the large amount of stochastic variability inherent in the model. By simulating the system 5000 times, one can collect the statistics of these variations and record them as functions of time. The dynamics of the mean levels of each species is shown by the solid, but somewhat jagged, black lines in Fig. 7. Furthermore, one can collect statistics on the number of monomers and tetramers at different points in time and plot the resulting histograms to show their marginal distributions as illustrated in Figs. 8 and 9. From these plots, it is noticeable that the deterministic reaction rate equations and the mean of the stochastic process are not equivalent for this process. This discrepancy arises from the non-linearity of the propensity functions for the the first and third reactions.

5.3 Normal Moment Closures

Above we have derived the differential equation for the mean of the process to be:

$$\frac{d}{dt}\mathbb{E}(\mathbf{X}) = \mathbf{S}\mathbb{E}\{\mathbf{w}(\mathbf{X})\} \tag{9}$$

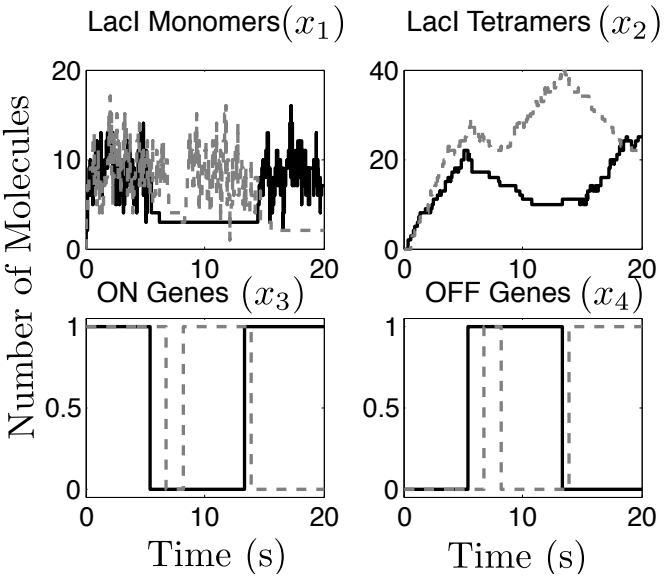

Figure 6: Results of two stochastic simulations (solid black, dashed gray) of the self-repressing LacI synthetic gene regulatory network. The top left panel corresponds to the populations of LacI monomers; the top right panel corresponds to the population of LacI tetramers; the bottom left corresponds to the population of ON genes; and the bottom right panel corresponds to the population of OFF genes.

In the case of linear propensity functions, then the average propensity function is simply the propensity function of the average population, and we could make the substitution:

$$\mathbf{S}\mathbb{E}\{\mathbf{w}(\mathbf{X})\} = \mathbf{S}\mathbf{w}(\mathbb{E}\{\mathbf{X}\}), \text{ for affine linear } \mathbf{w}(\mathbf{X}).$$

However, when the propensity function are non-linear, this substitution is incorrect, and in our case we have:

$$\mathbb{E}\{\mathbf{w}(\mathbf{X})\} = \mathbb{E}\left\{\begin{bmatrix} k_1\binom{x_1}{4} \\ k_2x_2 \\ k_3x_3x_2 \\ k_4x_4 \\ k_5x_3 \\ k_6x_1 \\ k_7x_2 \end{bmatrix}\right\} = \begin{bmatrix} k_1/24\left(\mathbb{E}\{x_1^4\} - 6\mathbb{E}\{x_1^3\} + 11\mathbb{E}\{x_1^2\} - 6\mathbb{E}\{x_1\}\right) \\ k_2\mathbb{E}\{x_2\} \\ k_3\mathbb{E}\{x_3x_2\} \\ k_4\mathbb{E}\{x_4\} \\ k_5\mathbb{E}\{x_3\} \\ k_6\mathbb{E}\{x_1\} \\ k_7\mathbb{E}\{x_2\} \end{bmatrix}.$$

Thus, we find that the expected values depend upon higher order moments, and the equations are not closed to a finite set. Similarly, the ODEs that describe the evolution of the second moments are given by:

$$\frac{d}{dt}\mathbb{E}\{\mathbf{X}\mathbf{X}^T\} = \mathbf{S}\mathbb{E}\{\mathbf{w}(\mathbf{X})\mathbf{X}^T\} + \mathbb{E}\{\mathbf{w}(\mathbf{X})\mathbf{X}^T\}^T\mathbf{S}^T + \mathbf{S}\left\{\text{diag}(\mathbb{E}\{\mathbf{w}(\mathbf{X})\})\right\}\mathbf{S}^T, \tag{10}$$

where the matrix $\mathbf{w}(\mathbf{x})\mathbf{x}^T$ is

$$\mathbf{w}(\mathbf{X})\mathbf{X}^T = \begin{bmatrix} k_1\binom{x_1}{4}x_1 & k_1\binom{x_1}{4}x_2 & k_1\binom{x_1}{4}x_3 & k_1\binom{x_1}{4}x_4 \\ k_2x_1x_2 & k_2x_2^2 & k_2x_2x_3 & k_2x_2x_4 \\ k_3x_1x_2x_3 & k_3x_2^2x_3 & k_3x_2x_3^2 & k_3x_2x_3x_4 \\ k_4x_1x_4 & k_4x_2x_4 & k_4x_3x_4 & k_4x_4^2 \\ k_5x_1x_3 & k_5x_2x_3 & k_5x_3^2 & k_5x_3x_4 \\ k_6x_1^2 & k_6x_1x_2 & k_6x_1x_3 & k_6x_1x_4 \\ k_7x_1x_2 & k_7x_2^2 & k_7x_2x_3 & k_7x_2x_4 \end{bmatrix},$$

In this case we see that the second moment also depends upon higher order moments. In particular the second moment of x_1 now depends upon the fifth uncentered moment of x_1. This relationship will continue for every higher moment such that the n^{th} moment will always depend upon the $(n+3)^{rd}$ order moment for this system.

If we make the assumption that the joint distribution of all species are given by a multivariate normal distribution, then we can use this relationship to close the moment equations. Perhaps the easiest way to find these relationships is to use the moment generating function approach. We define the MGF as:

$$M_{\mathbf{x}}(\mathbf{t}) = \exp\left(\mu^T\mathbf{t} + 1/2\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\right),$$

where the vectors are defined as:

$$\begin{aligned} \mu &= \mathbb{E}\{\mathbf{x}\}, \\ \mathbf{\Sigma} &= \mathbb{E}\{(\mathbf{x}-\mu)(\mathbf{X}-\mu)^T\} \\ &= \mathbb{E}\{\mathbf{X}\mathbf{X}^T\} - \mathbb{E}\{\mathbf{X}\}\mathbb{E}\{\mathbf{x}^T\}, \text{ and} \\ \mathbf{t} &= [t_1, t_2, t_3, t_4]^T. \end{aligned}$$

With this definition, one can write any uncentered moment in terms of μ and $\mathbf{\Sigma}$ as follows:

$$\mathbb{E}\{x_1^{n_1}\dots x_4^{n_4}\} = \frac{d^{n_1+\dots+n_4}}{dx_1^{n_1}\dots dx_4^{n_4}}M_x(\mathbf{t})\bigg|_{\mathbf{t}=\mathbf{0}}.$$

For example, the fifth uncentered moment of x_1 is given by:

$$\begin{aligned} \mathbb{E}\{x_1^5\} &= \frac{d^5}{dx_1^5}M_x(\mathbf{t})\bigg|_{\mathbf{t}=\mathbf{0}} \\ &= 15\mathbb{E}\{x_1\}\mathbb{E}\{x_1^2\}^2 - 20\mathbb{E}\{x_1\}^3\mathbb{E}\{x_1^2\} + 6\mathbb{E}\{x_1\}^5. \end{aligned}$$

Such an expression can be found for each moment of order three or higher in Eqns. 9 and 10. As a result the approximated distribution is fully described in terms of the first and second moments, which are our new set of fourteen dynamic variables:

$$\begin{aligned}
&\mathbb{E}\{x_1\}, \mathbb{E}\{x_2\}, \mathbb{E}\{x_3\}, \mathbb{E}\{x_4\}, \\
&\mathbb{E}\{x_1^2\}\mathbb{E}\{x_1x_2\}, \mathbb{E}\{x_1x_3\}, \mathbb{E}\{x_1x_4\}, \\
&\mathbb{E}\{x_2^2\}, \mathbb{E}\{x_2x_3\}, \mathbb{E}\{x_2x_4\}\}, \\
&\mathbb{E}\{x_3^2\}, \mathbb{E}\{x_3x_4\}, \mathbb{E}\{x_4^2\}\}.
\end{aligned} \tag{11}$$

We note that because there is only a single gene then x_3 and x_4 are mutually exclusive and take values of either zero or one. As a result, we can specify algebraic constraints on the last three of the moments listed in (11) as:

$$\begin{aligned}
\mathbb{E}\{x_3^2\} &= \mathbb{E}\{x_3\}, \\
\mathbb{E}\{x_4^2\} &= \mathbb{E}\{x_4\}, \\
\mathbb{E}\{x_3x_4\} &= 0
\end{aligned}$$

and thus we are left with only eleven ordinary differential equations.

We have solved the non-linear ODE's resulting from the moment closure, and the results for the mean values of each species are represented by the gray dashed lines in Fig. 7. From the figure, we see that for this case, the use of the coupled first and second moments results in a much better approximation of the of the mean behavior than did the deterministic reaction rate equation (compare solid and dashed gray lines in Fig. 7).

By including some description of the second uncentered moment of the process, the moment closure does a much better job of capturing the mean behavior of the process as can be seen by Fig. 7. Furthermore, closer examination reveals that the second moment for the population of monomers is also well captured by this approximation as is seen in Fig. 8. However, it is clear that the actual distributions are not Gaussian, and truncating away the higher order moments has introduced significant errors. This can be seen first in the monomer distributions at $t = 10s$, where the actual distribution appears to be almost bimodal. An even worse approximation is obtained for the tetramer distribution as is shown in Fig. 9, where the solution of the moment closure equations actually produces a physically unrealizable result of negative variance for the tetramer distribution. This failure is not unexpected due to the fact that the dynamics of the tetramer population depend strongly on the approximated high order moments of the monomer population.

5.4 FSP Analysis

In general the master equation can be written in the form $\mathbf{P}(t) = \mathbf{A}\mathbf{P}(t)$, where the infinitesimal generator $\mathbf{A}$ is defined as:

$$A_{i_2 i_1} = \begin{cases} -\sum_{k=1}^{M} w_k(\mathbf{x}_{i_1}) & \text{for } i_1 = i_2 \\ -w_k(\mathbf{x}_{i_1}) & \text{for } \mathbf{x}_{i_2} = \mathbf{x}_{i_1} + \mathbf{s}_k \\ 0 & \text{otherwise} \end{cases}$$

However, in order for this notation to make sense, one first has to define the enumeration of all the possible states $\{\mathbf{x}\}$. Based upon a few runs of the SSA, we can restrict our attention to a finite region of the state space ($x_1 \leq N_1 = 30$ and $x_2 \leq N_2 = 55$), then we can use the following scheme.

$$i(\mathbf{x}) = x_4(N_1 + 1)(N_2 + 1) + x_1(N_2 + 1) + x_2 + 1.$$

Note that we can make this enumeration depends only on x_1, x_2 and x_4 due to the fact that x_3 and x_4 are mutually exclusive and $x_3 = 1 - x_4$.

The FSP analysis has been conducted and the black dotted lines in Fig. 7 show the mean value of each of the four species as functions of time. With the chosen projection, the total one norm error in the computed probability distribution is guaranteed to be 4.8×10^{-5} or less at every instant in time. As such, the FSP solution makes a good basis upon which to compare the other solution schemes. With the FSP solution we can also determine not just the mean but the entire probability distribution at each time point, and the marginal distributions of the monomers (x_1) and the tetramers (x_2) are shown at times t=$\{$0.5,1,5,10$\}$s in Figs. 8 and 9.

6 Acknowledgments

The authors acknowledge support by the National Science Foundation under grants ECCS-0835847 and ECCS-0802008, the Institute for Collaborative Biotechnologies through Grant DAAD19-03-D-0004 from the US Army Research Office, and Los Alamos LDRD funding.

References

[1] H. El-Samad and M. Khammash. Stochastic stability and its applications to the study of gene regulatory networks. In *Proceedings of the 43rd IEEE Conference on Decision and Control*, 2004.

[2] H. El-Samad and M. Khammash. Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks. *Biophysical Journal*, 90:3749–3761, 2006.

[3] Johan Elf and Mns Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.*, 13:2475–2484, 2003.

[4] M. Elowitz, A. Levine, E. Siggia, and P. Swain. Stochastic gene expression in a single cell. *Nature*, 297(5584):1183–1186, 2002.

[5] M.B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.

[6] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes Characterization and Convergence*. Wiley Series in Probability and Statistics, 1986.

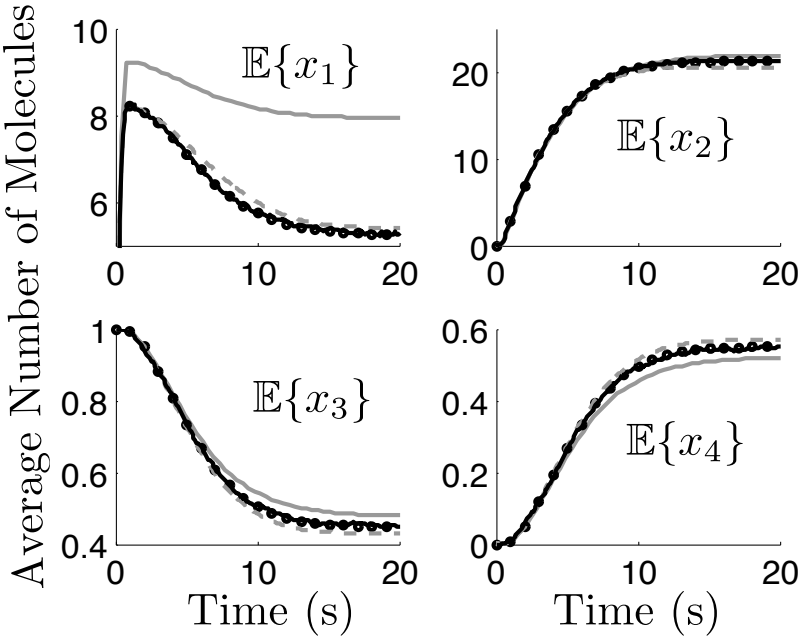

Figure 7: Dynamics of the mean values of $\mathbf{x}$ as found using different solution schemes. The solid gray lines correspond to the solution of the deterministic reaction rate equations. The dashed gray lines correspond to the solution using moment closure based upon the assumption of a multivariated Gaussian distribution. The jagged black lines correspond to the solution of 5000 stochastic simulations. The dotted lines correspond to the solution with the Finite State Projection approach.

[7] T.S. Gardner, C.R. Cantor, and J.J. Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403:339–342, 2000.

[8] Dan T. Gillespie. The chemical langevin and fokker-planck equations for the reversible isomerization reaction. *Journal of Physical Chemistry*, 106:5063–5071, 2002.

[9] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. of Computational Physics*, 22:403–434, 1976.

[10] Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115:1716–1733, 2001.

[11] C. A. Gomez-Uribe and G. C. Verghese. Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations. *J. of Chemical Physics*, 126(2):024109–024109–12, 2007.

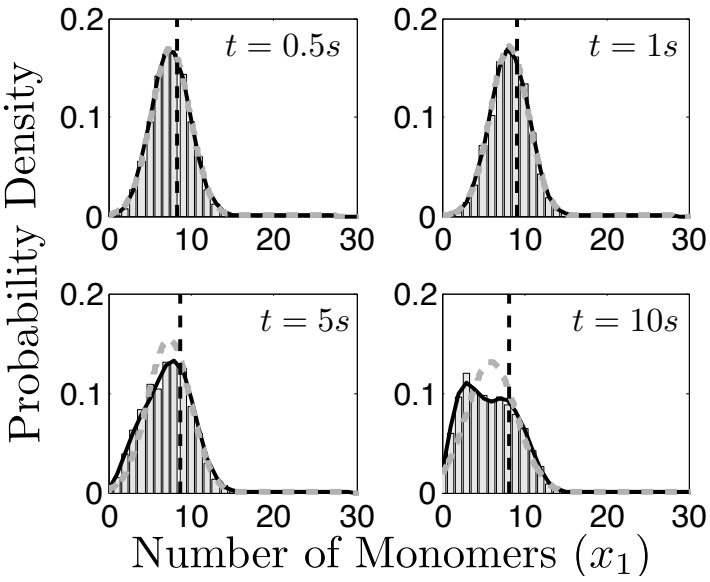

Figure 8: Probability distribution for the number of LacI monomers (x_1) at different points in time. The grey histograms have been found using 5000 stochastic simulations. The solid black lines correspond to the FSP solution. The dashed black lines shows the prediction of the mean using the deterministic reaction rate equations, and the dashed gray lines shows the results of the moment closure approach with an assumption of a normal distribution.

[12] J. Hasty, D. McMillen, and J.J. Collins. Engineered gene circuits. *Nature*, 420(6912):224–230, 2002.

[13] Joo Pedro Hespanha. Polynomial stochastic hybrid systems. In Manfred Morari and Lothar Thiele, editors, *Hybrid Systems: Computation and Control*, number 3414 in Lect. Notes in Comput. Science, pages 322–338. Springer-Verlag, Berlin, March 2005.

[14] F.J. Isaacs, J. Hasty, C.R. Cantor, and J.J. Collins. Prediction and measurement of an autoregulatory genetic module. *Proc. Natl. Acad. Sci. USA*, 100:7714–7719, 2003.

[15] M. J. Keeling. Multiplicative moments and measures of persistence in ecology. *J. of Theoretical Biology*, 205:269–281, 2000.

[16] M. Khammash. *Control Theory in Systems Biology, Chapter 2.* MIT Press, Cambridge, MA, 2009.

[17] Mustafa Khammash and Hana El-Samad. Stochastic modeling and analysis of genetic networks. In *Proceedings of the 44th IEEE Conference on Decision and Control and 2005 European Control Conference*, pages 2320– 2325, 2005.

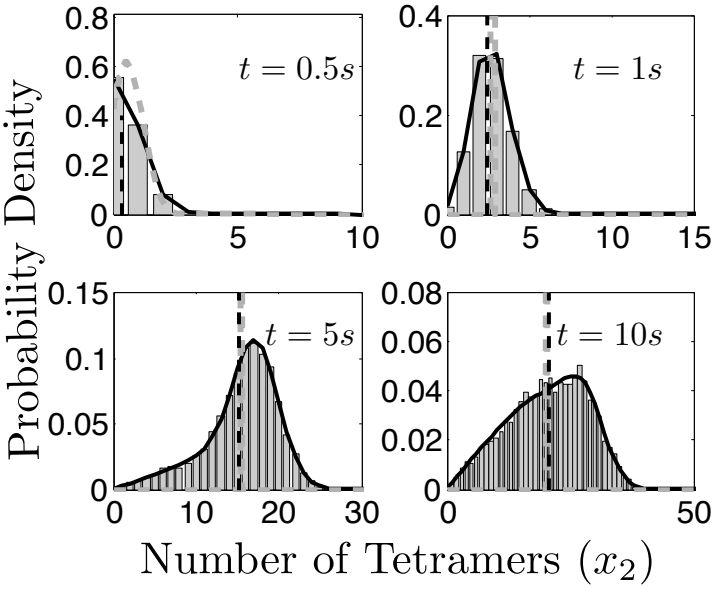

Figure 9: Probability distribution for the number of LacI tetramers (x_2) at different points in time. The grey histograms have been found using 5000 stochastic simulations. The solid black lines correspond to the FSP solution. The dashed black lines show the prediction of the mean using the deterministic reaction rate equations, and the dashed gray line shows the results of the moment closure approach with an assumption of a normal distribution.

[18] Thomas G. Kurtz. Strong approximation theorems for density dependent markov chains. *Stochastic Processes and their Applications*, 6:223–240, 1978.

[19] Harley H. McAdams and Adam Arkin. It's a noisy business! genetic regulation at the nanomolar scale. *Trends in Genetics*, 15(2):65–69, February 1999.

[20] HarleyH. McAdams and Adam Arkin. Stochastic mechanisms in geneexpression. *Proc. of the National Academy of Sciences U.S.A*, 94(3):814–819, 1997.

[21] B. Munsky. *The Finite State Projection Approach for the Solution of the Master Equation and its Application to Stochastic Gene Regulatory Networks.* PhD thesis, Univ. of California, Santa Barbara, Santa Barbara, 2008.

[22] B. Munsky, Hernday, D. Low, and Khammash. Stochastic modeling of the pap pili epigenetic switch. In *Foundations of Systems Biology in Engineering*, August 2005.

[23] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *Journal of Chemical Physics*, 124:044104, 2006.

[24] B. Munsky and M. Khammash. The finite state projection approach for the analysis of stochastic noise in gene networks. *IEEE Trans. on Automat. Contr.*, 53:201–214, January 2008.

[25] B. Munsky, B. Trinh, and M. Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5(318), 2009.

[26] Brian Munsky and Mustafa Khammash. Using noise transmission properties to identify stochastic gene regulatory networks. In *CDC08*, December 2008.

[27] I. Nasell. An extension of the moment closure method. *Theoretical Population Biology*, 64:233–239, 2003.

[28] I. Nasell. Moment closure and the stochastic logistic model. *Theoretical Population Biology*, 63:159–168, 2003.

[29] J. Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, 2004.

[30] Johan Paulsson, Otto Berg, and Mns Ehrenberg. Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97:7148–7153, 2000.

[31] Juan M. Pedraza and Alexander van Oudenaarden. Noise propoagation in gene networks. *Science*, 307(5717):1965 – 1969, March 2005.

[32] S. Peles, B. Munsky, and M. Khammash. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *Journal of Chemical Physics*, 20:204104, November 2006.

[33] N. Rosenfeld, M. Elowitz, and U. Alon. Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.*, 323:785–793, 2002.

[34] Abhyudai Singh and Joo Pedro Hespanha. A derivative matching approach to moment closure for the stochastic logistic model. *Bulletin of Mathematical Biology*, 69:1909–1025, 2007.

[35] P. Swain, M. Elowitz, and E. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences, USA*, 99(20):12795–12800, 2002.

[36] P.S. Swain. Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J. Mol. Biol.*, 344:965–976, 2004.

[37] M. Thattai and A. VanOudenaarden. Intrinsic noise in gene regulatory networks. *PNAS*, 98:8614–8619, 2001.

[38] M. Thattai and A. VanOudenaarden. Attenuation of noise in ultrasensitive signaling cascades. *Biophys. J.*, 82:2943–2950, 2002.

[39] R. Tomioka, H. Kimura, T.J. Koboyashi, and K. Aihara. Multivariate analysis of noise in genetic regulatory networks. *J. Theor. Biol.*, 229(3):501–521, 2004.

[40] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science, 2001.

[41] P. Whittle. On the use of the normal approximation in the treatment of stochastic processes. *J. Royal Statist. Soc., Ser. B*, 19:268–281, 1957.

[42] Mitsumasa Yoda, Tomohiro Ushikubo, Wataru Inoue, and Masaki Sasai. Roles of noise in single and coupled multiple genetic oscillators. *J. Chem. Phys.*, 126:115101, 2007.