LA-UR- 12-01601

Approved for public release;
distribution is unlimited.

Title:	Adaptive Coarse-Graining for Transient and Quasi-Equilibrium Analyses of Stochastic Gene Regulation
Author(s):	Brian Munsky (CCS3 and CNLS) Jose Juan Tapia (University of Pittsburgh) James Faeder (University of Pittsburgh)
Intended for:	51st IEEE Conference on Decision and Control Maui, Hawaii, USA December 9-13, 2012

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

UNCLASSIFIED DESIGNATION: BIOSCI

Abstract: Single-cell populations of genes, RNA and proteins are often described by continuous-time, discrete-state Markov processes. As such, the time-varying probability distributions of these systems evolve according to the large or infinite dimensional linear ordinary differential equation known as the (chemical) master equation (ME). Unfortunately, the large dimension of the ME means that numerical integration or stochastic simulation is often impossible or time consuming. To enable useful integration of computational and experimen- tal studies, new approximations are needed to improve this efficiency. In this paper, we introduce new methods to project the full ME onto a lower dimensional space, while retaining the transient and equilibrium statistics of the original process. First, we investigate three complementary sets of simple coarse- graining rules: (i) We use the previously described finite state projection approach to remove unlikely states from the Markov process; (ii) We modify an existing coarse-graining approach that reduces the dimension, but captures the equilibrium distri- bution of the process; and (iii) We introduce small correction terms to the time-scales of the reduced process in order to capture the transient dynamics of the original system. Next, we explore three different iterative algorithms that passively or actively adapt the projection resolution and thereby improve both accuracy and efficiency of the ME solution. We test the resulting projection rules and refinement strategies on a number of one-, two-, three- and four- species gene regulatory processes, and we select the most efficient and most accurate combination of coarse-graining rules and refinement strategies. Finally, we apply the final adaptive, coarse-grained ME solution to simultaneously adjust system parameters and approximation accuracy while fitting a stochastic model to simulated gene regulation data.

Adaptive Coarse-Graining for Transient and Quasi-Equilibrium Analyses of Stochastic Gene Regulation

José Juan Tapia,
Department of Computational and Systems Biology,
University of Pittsburgh,
Pittsburgh, PA 15260, USA
jjtapia@pitt.edu

James R. Faeder,
Department of Computational and Systems Biology.
University of Pittsburgh
Pittsburgh, PA 15260, USA
faeder@pitt.edu

Brian Munsky,*
Information Sciences Group,
The Center For Nonlinear Studies,
Los Alamos National Laboratory
Los Alamos, NM 87545
munsky@lanl.gov

Abstract— **Intracellular populations of genes, RNA and proteins are often described by continuous-time, discrete-state Markov processes. As such, the time-varying probability distributions of these systems evolve according to the large or infinite dimensional linear ordinary differential equation known as the chemical master equation (CME). Unfortunately, the large dimension of the CME means that numerical integration and stochastic simulation are often impossible or time consuming. To enable useful integration of computational and experimental studies, new approximations are needed to improve this efficiency. In this paper, we introduce new methods to project the full CME onto a lower dimensional space, while retaining the transient and equilibrium statistics of the original process. First, we investigate three complementary sets of simple coarse-graining rules: (i) We use the previously described finite state projection approach to remove unlikely states from the Markov process; (ii) We modify an existing coarse-graining approach in order to reduce the system dimension while capturing the equilibrium distribution of the process; and (iii) We introduce small correction terms to the time-scales of the reduced process in order to capture the transient dynamics of the original system. Next, we explore four different iterative algorithms that automatically adapt the projection resolution and thereby improve both accuracy and efficiency of the CME solution. We test the resulting projection rules and refinement strategies on a number of one-, two-, and three-species gene regulatory processes, and we select the most efficient and most accurate combination of coarse-graining rules and refinement strategies.**

I. Introduction

A system of reacting chemicals in a homogeneous environment can be described as a Markov chain where each state represents a discrete reactant population a given time. For such a Markov chain, one can define a probability density vector, which evolves according to a linear ordinary differential equation known as the chemical master equation (CME) [1]. Integrating the CME is often computationally intractable, especially when there are more than a few chemical species. Kinetic Monte Carlo (KMC) methods, such as the Gillespie stochastic simulation algorithm [2], have been developed to reconstruct the system probability distribution from sample trajectories. However, the speed of convergence of KMC-based solutions of the CME is slow–an n-fold improvement the precision requires n^2 times as many simulations.

*To whom correspondence should be addressed.

In many systems biology problems, a few representative trajectories are sufficient to elucidate the stochastic system's behavior. For such qualitative studies there is no real need to solve the CME, and simulation is sufficient. However, with the increased availability of single-cell and single-molecule data, it has become possible to measure the distributions of molecular populations [3], [4], [5], [6], [7], [8], [9]. With such data, it is now possible to identify stochastic gene regulatory models from the variable distributions of experimental systems [10], [11]. Such identification procedures require that CME solutions be precise enough to capture all features of the experimental data, yet fast enough to be solved for many different parameter combinations ($\gg 10^5$ in many parameter searches).

To improve parameter identification studies, we need to (i) increase the class of problems for which CME integration is feasible and (ii) increase the efficiency of this integration. To accomplish these objectives, several projection-based methods have been proposed to reduce the number of equations that must be integrated, including finite state projection approaches [12], [13], Krylov subspace methods [14], and time-scale partitioning techniques [15], [16]. One type of projection-based reduction of particular interest are the so-called sparse-gridding (or interpolation-based reduction) approaches [17], [13]. In these approaches, the probabilities of adjacent states in the natural configuration space are interpolated from among fewer points on a sparser grid.

The two key tasks of the sparse-grid CME reduction are to choose (i) the shape of the interpolation function (*i*.e., how the exact solution is assumed to relate to the approximate solution on a coarser mesh), and (ii) the coarseness of this actual mesh. Both selections introduce a tradeoff between efficiency and accuracy–choices that are broadly similar to those of using a variable step size and approximation order when numerically integrating a differential equation. In this paper, we introduce a few different interpolation shape functions and adaptive grid-refinement strategies, and we compare them in terms of their ability to reduce the dimension of the CME while retaining integration accuracy.

The remainder of this paper is organized as follows: In Section II, we introduce the general formulation of the CME, and we discuss a few approaches to reducing its dimension, such as the finite state projection approach (II-A) and zeroth-

order coarse-grid projections (II-B). Next in IV, we discuss three candidate approaches to adapt the coarse grid while solving the CME. In Section V, we describe a set of different test systems and evaluate the different reduction approaches. Finally, in Section VI, we summarize the performance of the different approaches and their applicability to larger chemical systems, and we make concluding remarks on the use of these tools for the identification of stochastic gene regulatory models.

II. Methods

A system of N interacting chemicals in a homogeneous environment can be described as a Markov chain where each state, $\mathbf{x_i} = [\xi_1, \ldots, \xi_S]_i^T \in \mathbb{X} \in \mathbb{N}_{\geq 0}^N$, represents specific integer populations of the N reactants. The reaction *propensity* (or stochastic reaction rate) functions, $w_u(\mathbf{x_i})\mathbf{dt}$, are the probabilities that each of the $u = (1 \ldots U)$ reactions will occur in the next infinitesimal time step dt. These reactions are transitions from one state, $\mathbf{x_i}$, to another state, $\mathbf{x_j} = \mathbf{x_i} + \mathbf{v_u}$. For such a Markov chain, the probability density vector, $\mathbf{P}(t) = [p_1(t), p_2(t), \ldots]^T$ represents the probability that $\mathbf{x}(t) = \mathbf{x}_i$ at time t. The probability density vector, $\mathbf{P}(t)$, evolves according to the linear ordinary differential equation, $\frac{d}{dt}\mathbf{P}(t) = \mathbf{A}(t)\mathbf{P}(t)$, which is known as the chemical master equation (CME). As discussed in the literature [12], [16], [13], the infinitesimal generator, $\mathbf{A} = [A_{ij}]$, can be specified as:

$$A_{ji} = \left\{ \begin{array}{ll} \sum_u w_u(\mathbf{x_i}) & \text{if } j = i \\ w_u(\mathbf{x_i}) & \text{if } \mathbf{x_j} = \mathbf{x_i} + \mathbf{v_u} \\ 0 & \text{otherwise.} \end{array} \right\} \quad (1)$$

Now that the CME is formulated, we turn to approaches to reduce and solve it.

A. Finite State Projection

The CME dimension can be extremely large or infinite, so approximations are needed. One such approximation, the finite state projection (FSP) approach [12], [13], selects a finite set states, $\mathbb{X}_J \in \mathbb{X}$, and aggregates the vast majority of remaining states, $\mathbb{X}_{J'}$, into one or more absorbing states. The end result is a finite-dimensional master equation,

$$\frac{d}{dt}\mathbf{P}^{\text{FSP}}(t) = \mathbf{A}_J(t)\mathbf{P}^{\text{FSP}}(t), \quad (2)$$

, which can be integrated numerically[1]. By keeping track of the probability lost to each absorbing sink, it is possible to compute the CME error and systematically expand the set $\mathbb{X}_J$ until that error satisfies a pre-specified threshold. For simple systems, such as two-species chemical reactions, efficient algorithms exist to define and expand the $\mathbb{X}_J$ (see [19] for a complete description of these algorithms). However, even after the application of the FSP, the dimension of Eq. 2 can be extremely large, such that numerical integration is prohibitively expensive. For these cases, additional reductions are necessary.

[1]All integration of master equations in this work is done in Matlab using Roger Sigje's *expokit* [18]

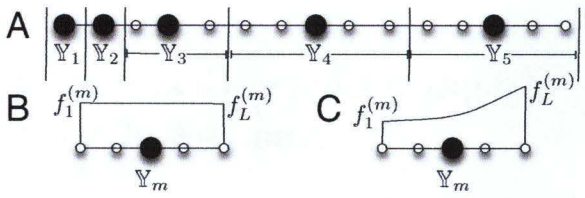

Fig. 1. Schematics of the one-species gridding approach and different possible shape functions. A) One-D grid for a zeroth-order interpolation, containing five elements $M = 5$. B) Zeroth order interpolation: one interpolation point per element; probability is equally distributed at all points in element (Eq. 6). C) Zeroth order, corrected interpolation: one interpolation point per element; probability is distributed in element according to Eq. 11.

B. Projection-Based Reductions of the ME

Even after application of the FSP, it is often useful to reduce the dimension of the CME through some type of projection operation [13]. In these projections, one assumes that $\mathbf{P}(t)$ can be approximated by a linear transformation of a lower dimensional vector,

$$\mathbf{P}(t) \approx \mathbf{\Phi}\mathbf{q}_\Phi(t), \quad (3)$$

where the matrix $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$ defines the projection operator (typically $M \ll N$). This operator defines the dynamics of the lower dimensional vector, $\mathbf{q}_\Phi(t)$:

$$\begin{aligned} \frac{d}{dt}\mathbf{q}_\Phi(t) &= \mathbf{\Phi}^{-L}\mathbf{A}\mathbf{\Phi}\mathbf{q}_\Phi(t), \\ \mathbf{q}_\Phi(0) &= \mathbf{\Phi}^{-L}\mathbf{P}(0), \end{aligned} \quad (4)$$

where $\mathbf{\Phi}^{-L}$ denotes the left inverse of $\mathbf{\Phi}$. This ODE can be solved for $\mathbf{q}_\Phi(t)$ and used to approximate $\mathbf{P}(t)$.

The intention of the interpolation reduction method is to reduce the complexity, yet retain accuracy, in the CME solution. For this, we seek to minimize the tradeoff:

$$\min_{\mathbf{\Phi}} \left\{ |\mathbf{P}(t_f) - \mathbf{\Phi}\mathbf{q}_\Phi(t_f)| + \lambda(\mathbf{\Phi}) \right\}, \quad (5)$$

where $\lambda(\mathbf{\Phi})$ penalizes computational complexity (or compute time) associated with the M-dimensional ODE in Eq. 4. Several different methods have been proposed to choose $\mathbf{\Phi}$, including Krylov subspace methods [14] and time-scale partitioning techniques [15], [16]. Here, we focus upon sparse-gridding or interpolation-based projections [17], [13].

III. Sparse-Grid Reductions of the CME

We focus on projection operations, that section $\mathbb{X}_J$ into a number of disjoint rectangular elements as illustrated in Fig. 1A. For each element, we restrict ourselves to zeroth-order interpolations approaches, which lump each element into a single coarse-grained state. Higher order interpolations can be formulated in a similar manner, but are not considered in this report. In what follows, we briefly describe these interpolation functions and the resulting definition of the projection operator $\mathbf{\Phi}$.

A. Zeroth-Order Interpolations

The grid in Fig. 1A defines a set of M disjoint rectangular regions, $\{\mathbb{Y}_m\}$, which cover the entire state space, $\mathbb{X}$. Each region contains exactly L_m configurations. The coarse

state $q_m(t)$ approximates the total probability of the points contained within the $\mathbb{Y}_m$. In the simplest coarse grid, the probabilities of points within a region are approximated as being equal, such that

$$p_i(t) \approx q_m(t)/L_{\mathrm{m}}, \text{ for } \mathbf{x_i} \in \mathbb{Y}_{\mathbf{m}}. \tag{6}$$

This leads to a very simple definition of $\mathbf{\Phi} = [\phi_{im}]$:

$$\phi_{im} = \begin{cases} 1/L_m & \text{if } \mathbf{x}_i \in \mathbb{Y}_m \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The left inverse, $\mathbf{\Phi}^{-L} = \nu_{mi}$ is simply:

$$\nu_{mi} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathbb{Y}_m \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The zeroth order interpolation approach is convenient for two reasons: First, it is very easy to specify the projection operator $\mathbf{\Phi}$ and to compute the reduced generator $\mathbf{A_\Phi} = \mathbf{\Phi}^{-L}\mathbf{A}\mathbf{\Phi}$; Second, the reduced system for $\mathbf{q}(t)$ is itself a Markov process and therefore retains many useful properties associated with such systems, such as $\sum_m q(t) = 1$ and $\sum_m dq(t)/dt = 0$ for all t. The two main disadvantages of the interpolation are (i) all states $\mathbf{x_i} \in \mathbb{Y}_{\mathbf{m}}$ are not equally probable as assumed in Eq. 6, and (ii) the time required to traverse from one end of an element to the next is non-instantaneous. To address these issues, we introduce a couple of simple corrections to the interpolation.

Corrections to the Zeroth-Order Interpolation Scheme.

Our first step toward improving the accuracy of the zeroth-order interpolation approach is to estimate how the probability density varies along each element. In particular, we wish to know the density at the borders between neighboring elements, as these probabilities define the transitions from one element to the next.

Consider the points in a single coarse element as illustrated in Fig. 1C. Let $e_l^{(m)}$ refer to the l^{th} specific state in the m^{th} element, and define $f_l^{(m)}$ as the probability of that specific state. To estimate $f_l^{(m)}$, we approximate the propensity functions, $w_m^{\pm}$ as being constant along each element, where $+/-$ denote reactions that increase or decrease the population, respectively.

Transitions within each small element are assumed to reach quasi-steady equilibrium much faster compared to the full process, such that

$$w_m^+ f_l^{(m)} = w_m^- f_{l+1}^{(m)}, \text{ or} \tag{9}$$

$$f_{l+1}^{(m)} = r_m f_l^{(m)}, \tag{10}$$

where $r_m = w_m^+/w_m^-$. Since the total probability of the m^{th} element is $q_m(t)$, the probability of each state can be solved for as:

$$f_l^{(m)} = q_m(t)\left(\frac{1-r_m}{1-r_m^{L_m}}\right) r_m^{l-1}. \tag{11}$$

In particular, the probability densities at the borders of each element are now approximated as:

$$\begin{bmatrix} f_1^{(m)}(t) \\ f_{L_m}^{(m)}(t) \end{bmatrix} = q_m(t)\left(\frac{1-r_m}{1-r_m^{L_m}}\right)\begin{bmatrix} 1 \\ r_m^{L_m-1} \end{bmatrix}. \tag{12}$$

Next, we take into consideration the time needed for the system to transition from the center of one element to the center of the next. The distance between the m^{th} and the $(m+1)^{\text{th}}$ centers is simply $(L_m + L_{m+1})/2$. The coarse-grained rates to go from the m^{th} to the $(m+1)^{\text{th}}$ or $(m-1)^{\text{th}}$ elements are then approximated as

$$\omega_m^+ = w_m^+ \frac{f_{L_m}^{(m)}(t)}{(L_m + L_{m+1})/2}, \text{ and}$$

$$\omega_m^- = w_m^- \frac{f_1^{(m)}(t)}{(L_m + L_{m-1})/2}. \tag{13}$$

The end result of correcting for the distribution of probability along each element and the time it requires to traverse each element, is that we have a reduced Markov process. For example, in Fig. 1, where the original process contained $N = 15$ states, the new one has only $M = 5$ states. Where the propensity function of the original process were w^+ and w^-, the new process has propensity functions ω_m^+ and ω_m^- as defined by Eq. 13.

Extending the corrected zeroth order coarse graining approach to additional species can be achieved simply by applying Eqs. 12 and 13 to each species.

IV. Adaptive Grid Selection Methods

Given a grid, we can use the interpolation functions from the previous section to project the CME down to a lower-dimensional space and approximate its solution. But how do we choose the grid to begin with? In this section, we focus on adaptive methods for grid selection. We begin with an initial grid space, where grid lines in each dimension are separated by 64 states. In other words, $L_m = 64$ for every m and every chemical species. This grid used to generate a reduced CME as described above, and the FSP approach is applied to select the important coarse states from the reduced system. Then, based upon the solution under the current grid, we divide or combine neighboring elements, being careful to ensure that all grid elements contain at least one state. In the following subsections, we describe three methods in detail: (i) passive refinement, (ii) probability-based refinement, and (iii) two-grid error control refinement. Each approach is described in the context of a one-species Markov Chain. For multiple dimensions, the grid refinement is done independently in each dimension.

A. Passive Refinement

In the passive grid-refinement strategy, we simply split every mesh element in each iteration. For example, for initial mesh lengths of $L_{m_0} = 64$ for $m = 1, \ldots, M_0$, the mesh lengths of the second iteration would be $L_{m_1} = 32$ for $m = 1, \ldots, M_1 (= 2M_0)$, and so on. For each successive grid definition, we can solve for the approximate probability distributions

$$\mathbf{P}_k(t) = \mathbf{\Phi}_k \mathbf{q}_k(t) \tag{14}$$

This process is continued until the one-norm difference in the probability distribution from one iteration to the next is

below a preset level of tolerance:

$$|\mathbf{P}_{k+1}(t) - \mathbf{P}_k(t)|_1 \leq tol, \tag{15}$$

or until every grid element consists of exactly one state.

B. Probability Concentration

The second refinement algorithm uses upper and lower thresholds on the approximated probabilities to refine the grid selection. Suppose that for the current refinement iteration, the solution of the reduced system is $\mathbf{q}(t_f)$, which has an average value of $\bar{q}$. Two refinement thresholds are specified: Δ_{max} and Δ_{min}. High probability elements, where

$$q_m \geq \Delta_{max} \cdot \bar{q}, \tag{16}$$

are subdivided into two disjoint elements. Adjacent low probability elements, where both

$$q_m + q_{m+1} \leq \Delta_{min} \cdot \bar{q}, \tag{17}$$

are merged together. Otherwise, the grid is left unaltered. For this approach, the refinement strategy is continued until no further refinements are possible or until Eq. 15 is satisfied between two successive iterations.

C. Two Grids

The "two grid" scheme is similar to a technique called "step doubling," which is often used for implementing error control in higher order Runge-Kutta methods. The underlying idea is as follows: Given a coarse grid, G_c, we define a finer grid G_f in the same manner as in the Section IV-A. These two grids are integrated separately to find $\mathbf{q}_f$ and $\mathbf{q}_c$. The two different solutions are then compared at the coarser level of detail, G_c, which provides a vector of local errors:

$$\mathbf{E}_m = |\mathbf{q}_{f_m} - \mathbf{q}_{c_m}| \tag{18}$$

where each $\mathbf{E}_m$ corresponds to the integration error of the less accurate of the two approximate solutions, $\mathbf{q}_c$, at grid point m.

If an entry of $\mathbf{E}_m$ is below threshold Δ_{max}, the corresponding mesh point is accepted, otherwise the element is divided into two. As before, the refinement strategy is continued until no further refinements are possible or until Eq. 15 is satisfied between G_c and G_f.

V. Examples

To illustrate the methods above, we apply them to a few simple gene regulatory network models. Specifically, we consider a general network, where N different species can activate or repress each one another by a complicated set of nonlinear production and degradation reactions. For each n^{th} species, the propensity function for a production reaction is given as:

$$w_n^+ = k_n^{(0)} + k_n^{(1)} \frac{\prod_{s=1}^{N} (1 + b_{sn} \xi_s^{\eta_{sn}})}{\prod_{s=1}^{N} (1 + a_{sn} \xi_s^{\eta_{sn}})}, \tag{19}$$

where ξ_s is the population of the s^{th} species; $k_n^{(0)}$ and $k_n^{(1)}$ are the basal and active production rates for species n; a_{sn} and b_{sn} are the rates at which species s represses or activates species n; and η_{sn} is the order of that activation/repression reaction. Furthermore, each species is assumed to decay as a first order reaction

$$w_n^- = \gamma_n \xi_n \tag{20}$$

With the proper choice of parameters, this general form encompasses many different models of stochastic gene regulation from the literature. Here, we will discuss three models in particular: (1) simple birth-death or Poisson process, (2) the genetic toggle switch, and (3) the three-species repressilator model.

A. Poisson Process

One of the simplest stochastic models of gene regulation is that of the simple birth-death process. In this process there is only a single species, and the production rate is simply: $w_i^+ = k$. Although very simple, the Poisson process is worth studying for two reasons. First, it matches the behavior of a large number of constitutively expressed genes in yeast [5]. Second, if the birth-death process begins with an initial Poisson distribution with mean $\mu(0)$, then it remains Poisson-distributed for all future times, and its mean evolves according to:

$$\frac{d\mu(t)}{dt} = k - \gamma\mu(t). \tag{21}$$

Since an exact solution is available, this system makes an ideal testbed upon which to evaluate different CME approximations.

To test our different methods, we first consider a 1-species "reaction," described by the Poisson process. The degradation rate is set to $\gamma = 1$ and production rate is set to, $k = 40$, corresponding to steady state mean of 40 proteins. The process starts at $\xi(0) = 0$ at $t = 0$. Fig. 2, plots the computed distributions at times $t = 0.5$ (left) and $t = 5$ (right) (in arbitrary units of $1/\gamma$). In Fig. 2, we consider three different mesh sizes from coarse (top row) to fine (bottom row). The different lines in the figures correspond to the exact in black and approximations using different zeroth-order interpolation functions in the various colors. In red circles, we plot the un-corrected zeroth-order interpolations. In green triangles, we correct for the shape of the distribution in each element according to Eq. 2, and in blue crosses, we also include the correction for the transition time across each element. As the mesh becomes finer, all three interpolation approximations converge to the exact solution (see Panel C). However, the rate of convergence is not equivalent. For long times ($t = 5$, right panels), the two approaches that correct for the shape of the distribution do a good job of matching the quasi-steady state distribution (Fig. 2B, right. Both green and blue give a good approximation), but the uncorrected zeroth-order approximation does more poorly (red circles give a poor approximation). At short transient times ($t = 0.5$, Fig. 2B, left), approximations that lack a correction for the transit times (e.g., the green triangles), evolve faster than the true system, but this issue is well-corrected by Eq. 13 (blue crosses). Considering its success in approximating both transient and quasi-stationary distributions, we use the

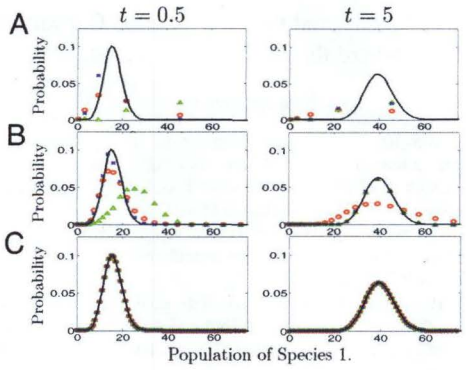

Fig. 2. Distributions of ξ in the Poisson process at times $t = 0.5$ (left) and $t = 5$ (right). The initial conditions i $\xi(0) = 0$. Distributions are computed using three different logarithmically distributed meshes with lengths round($\{h^{i-1}\}$) for $i = \{1, 2, ...\}$. (A) Coarse mesh, $h = 2$. (B) Moderate mesh, $h = 1.1$. (C) Fine mesh, $h = 1.01$. In each plot, the exact solution is shown with a solid black line; the zeroth-order mesh is shown in red circles; the zeroth-order mesh with distribution shape correction is shown with green triangles; and the zeroth-order mesh with distribution shape and transit time corrections are shown with blue crosses.

corrected zeroth-order interpolation approach (blue crosses) from this point forward.

B. Toggle Switch

The second model that we consider is the genetic toggle switch [20], which consists of two genes, where each gene is negatively regulated by the other. In the stochastic model of the toggle switch, the production rates of the two species is given by:

$$w_n^+ = k_n^{(0)} + k_n^{(1)} \frac{1}{(1 + a_{sn} \xi_s^{\eta_{sn}})}, \tag{22}$$

for $(n, s) = (1, 2)$ or $(2, 1)$, and $k_n^{(0)}$ are the basal production rates and $k_n^{(1)}$ are the active production rates for each species. For our simulations, we have chosen the parameters:

$$\left\{ \; k_n^{(0)} = 1 \quad k_n^{(1)} = 50 \quad a_{sn} = 5 \quad \eta_{sn} = 2\gamma = 1 \; \right\} \tag{23}$$

Fig. 3 shows the marginal distribution of species 1 at short and long times of $t = 1, 10$ units of γ.[2] Although no exact solution exists for this system, we have previously shown that the FSP approach provides an arbitrarily close approximation to the exact solution [11]. With this, we can now evaluate how close the approximate solution is to the exact solution for different mesh refinement strategies.

For the 2-state toggle model and a 2-state Poisson model (a simple extension of the process in Section V-A, where each species undergoes an independent birth-death process), we start with a coarse mesh in which every element is 64 states long. We then refine it according to the strategies discussed in Section IV. Fig. 4 plots the 1-norm error in the distribution at time, $t = 10$ versus the number of states in the reduced Markov chain and for each of the different mesh refinement strategies. Fig. 4A applies only

[2]Note, we have chosen the system to be symmetric, so the marginal distributions of the two species are equal.

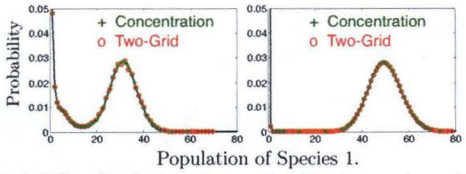

Fig. 3. Probability distributions for species 1 in the toggle model at $t = 1$ (left) and $t = 10$ (right). Initial conditions are $\xi_1(0) = \xi_2(0) = 0$. Green crosses correspond to the final approximation found with the concentration refinement algorithm (Section IV-B), and red circles correspond to the distribution found with the two grid approach (Section IV-C)

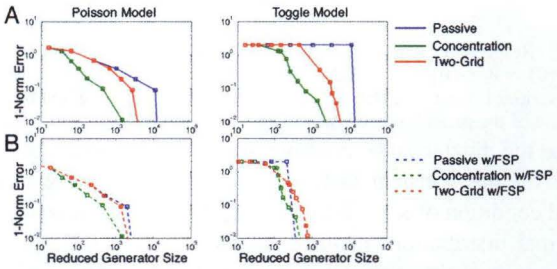

Fig. 4. Comparison of the number of bins versus the total probability distribution error for the 2-state Poisson model (left) and the 2-state toggle model (right). Different grid refinement strategies have been considered: passive refinement (Section IV-A, blue), probability concentration refinement (Section IV-B, green), and two-grid refinement (IV-C, red). (A) Without FSP reduction. (B) With FSP reduction.

the mesh refinement strategies, whereas Fig. 4B applies first the meshing strategy, followed by the finite state projection approach. For both systems, the adaptive grid approaches provide a significant improvement over the passive approach (compare red and green lines to blue line). For example, for the toggle model in Fig. 4A(right) the concentration-based adaptive grid refinement strategy reduces the number of states by an order of magnitude while maintaining the same level of accuracy. Applying the FSP reduction on top of the grid refinement strategy yields another large reduction in the system dimension, yet in most cases the adaptive refinement strategies retain an advantage over the passive approach.

C. Repressilator

The third system that we consider is a three species repressilator [21], which consists of three chemical species that regulate each other through repression in a sequential feedback loop. It can be considered as a higher dimensional version of the toggle switch, but can produce significantly different behaviors. The production rates for each of the three species is:

$$\begin{aligned} w_1^+ &= k^{(0)} + k^{(1)} \frac{1}{(1 + a\xi_2^{\eta})} \\ w_2^+ &= k^{(0)} + k^{(1)} \frac{1}{(1 + a\xi_3^{\eta})} \\ w_3^+ &= k^{(0)} + k^{(1)} \frac{1}{(1 + a\xi_1^{\eta})} \end{aligned} \tag{24}$$

where

$$\left\{ \; k^{(0)} = 0 \quad k^{(1)} = 25 \quad a = 5 \quad \eta = 6 \quad \gamma = 1. \; \right\} \tag{25}$$

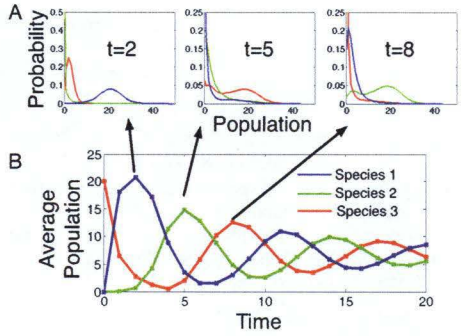

Fig. 5. Response of repressilator model to an initial condition of $\xi_1(0) = 20, \xi_2(0) = 0, \xi_3(0) = 0$. (A) Distributions of each of the three species at times various times $t = 2$ (left), $t = 5$ (middle), and $t = 8$ (right). (B) The response of the mean levels of each of the three species.

Using the final coarse graining method, we were able to solve for the transient distribution dynamics starting at an initial condition of $\mathbf{x} = [20, 0, 0]$. Fig. 5A shows the resulting marginal distributions at three points in time $t \in \{2, 5, 8\}$, where each of the species has become the dominant species. Although each individual system continues to oscillate indefinitely, different cells lose their synchronicity, and the distributions (and therefore the average populations) eventually converge to steady state as shown in Fig. 5B.

VI. Discussion

The dynamics of genes, RNA molecules and proteins in single cells are often described by probability distributions, which evolve according to the large set of ordinary differential equations, known as the chemical master equation. In many cases, the dimension of the CME is too large for it to be solved efficiently, and reductions to the CME are necessary. In this paper, we introduced a new approach to achieve such a reduction. These reductions are similar to finite element approaches, in that we assume that the probability density of states within small regions of the configuration space (e.g., elements) follow simple distribution shapes. We presented two zeroth-order interpolation schemes, which account for the variation of probability over each element and for the amount of time that is taken to transit from one side of the element to the next. We also suggested three possible approaches for refining the grid on which the approximation is made.

Combining the corrected zeroth-order interpolation shapes with the adaptive grid refinement strategies allowed us to achieve noticeable reductions in the size of the CME for a relatively small loss of accuracy. Considering that the computational effort to integrate the CME scales with $O(M^3)$ [22], this represents a potential computational savings of many orders of magnitude. Such improvements will eventually enable us to solve the CME for more complicated systems and will enable faster CME solutions for use in parameter idenfication studies.

Acknowledgment

This work was supported in part by NSF Expeditions in Computing Grant (award 0926181). JJT also acknowledges travel support from the Fifth q-bio Summer School (Center for Nonlinear Studies and the New Mexico Consortium, Los Alamos, NM), where this work was initiated.

References

[1] D. T. Gillespie, "A Rigorous Derivation of the Chemical Master Equation," *Physica A*, vol. 188, pp. 404–425, 1992.

[2] ——, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, pp. 2340–2361, 1977.

[3] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA Synthesis in Mammalian Cells," *PLoS Biol*, vol. 4, no. 10, p. e309, Sep. 2006.

[4] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, "Imaging individual mRNA molecules using multiple singly labeled probes," *Nature Publishing Group*, vol. 5, no. 10, pp. 877–879, Sep. 2008.

[5] D. Zenklusen, D. R. Larson, and R. H. Singer, "Single-RNA counting reveals alternative modes of gene expression in yeast," *Nature Structural and Molecular Biology*, vol. 15, no. 12, pp. 1263–1271, Nov. 2008.

[6] Y. Taniguchi, P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells," *Science*, vol. 329, no. 5991, pp. 533–538, Jul. 2010.

[7] L.-h. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding, "General properties of transcriptional time series in Escherichia coli," *Nature genetics*, vol. 43, no. 6, pp. 554–560, May 2011.

[8] S. J. Gandhi, D. Zenklusen, T. Lionnet, and R. H. Singer, "Transcription of functionally related constitutive genes is not coordinated," *Nature Structural & Molecular Biology*, vol. 18, no. 1, pp. 27–34, Dec. 2010.

[9] J. Stewart-Ornstein, J. S. Weissman, and H. El-Samad, "Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae," *Molecular Cell*, vol. 45, no. 4, pp. 483–493, Feb. 2012.

[10] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters." *Molecular systems biology*, vol. 5, p. 318, 2009.

[11] B. Munsky and M. Khammash, "Identification from stochastic cell-to-cell variation: a genetic switch case study." *IET Syst. Biol.*, vol. 4, no. 6, pp. 356–366, Nov. 2010.

[12] ——, "The finite state projection algorithm for the solution of the chemical master equation," *The Journal of chemical physics*, vol. 124, p. 044104, 2006.

[13] ——, "The finite state projection approach for the analysis of stochastic noise in gene networks," *Automatic Control, IEEE Transactions on*, vol. 53, no. Special Issue, pp. 201–214, 2008.

[14] K. Burrage, F. Macnamara, and B. Sidje, "A krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems." in *Mam 2006: Markov Anniversary Meeting*. C&M Online Media, Inc., 2006, p. 21.

[15] E. Haseltine and J. Rawlings, "Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics," *The Journal of Chemical Physics*, vol. 117, p. 6959, 2002.

[16] S. Peles, B. Munsky, and M. Khammash, "Reduction and solution of the chemical master equation using time scale separation and finite state projection." *The Journal of chemical physics*, vol. 125, no. 20, p. 204104, Nov. 2006.

[17] M. Hegland, C. Burden, L. Santoso, S. MacNamara, and H. Booth, "A solver for the stochastic master equation applied to gene regulatory networks," *Journal of Computational and Applied Mathematics*, vol. 205, no. 2, pp. 708–724, Aug. 2007.

[18] R. B. Sidje, "EXPOKIT: Software package for computing matrix exponentials," vol. 24, no. 1, pp. 130–156, Mar. 1998. [Online]. Available: http://doi.acm.org/10.1145/285861.285868

[19] B. Munsky, "Modeling cellular variability," *in Quantitative Biology From Molecular to Cellular Systems*, 2012.

[20] T. Gardner, C. Cantor, and J. Collins, "Construction of a genetic toggle switch inescherichia coli," *Nature*, vol. 403, pp. 339–342, 2000.

[21] M. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators." *Nature*, vol. 403, no. 6767, pp. 335–338, 2000.

[22] C. Moler and C. V. Loan, "Nineteen dubious ways to compute the exponential of a matrix," *SIAM Review*, vol. 20, pp. 801–836, 1978.