

Diagnosis of Rare Events in Systems Described by the Chemical Master Equation

David Thorsley

Department of Electrical Engineering, University of Washington
thorsley@u.washington.edu

Abstract. Biochemical processes inside the cell are often modeled at the mesoscopic scale as continuous-time Markov processes whose probability distributions evolve over time according to the chemical master equation. We consider the problem of determining the *a posteriori* probability that a particular rare event of interest has occurred in a chemical process given a record of observations of that process. Expressions for this *a posteriori* probability are developed for the case where the record of observations is continuous and for the case where it is intermittent, and the dynamics of the *a posteriori* probabilities are expressed as hybrid systems. The approach is demonstrated on a stochastic model of gene expression inside the cell.

1 Introduction

The biochemical processes that make up life at the cellular level are inherently stochastic [1]. Deterministic methods of modeling and analyzing chemical reactions, such as ordinary differential equation models, are based on the assumption that there is a large population of every species present in the reaction chamber. In biochemical systems such as gene regulatory networks, this assumption does not hold; the populations of species such as proteins, RNA strands, and genes may be very small. As a result of these small populations, biochemical processes can be intrinsically noisy; this noisy behavior has been observed experimentally in single-cell studies [2] and quantitatively analyzed [3], [4].

Under a standard set of physical assumptions, the behavior of a small-volume system can be described exactly by a continuous-time Markov process. Each state of the Markov process is an p -dimensional vector (where p is the number of distinct species in the reaction) whose elements are the population of each species in the reaction. Each transition in the process corresponds with the occurrence of one of the possible reactions.

When modeling biochemical processes as Markov processes, there are many different chemical species in the reaction chamber and the state space is thus very high-dimensional. As a result, stochastic biochemical processes are often studied using simulation techniques such as the Gillespie algorithm [5] and its extensions [6]. However, in order to use simulation methods to reliably estimate the probability distributions of aspects of a system's behavior, it is necessary

to take a very large number of simulations. This problem is compounded when the behaviors of biological interest are rare events, such as the state of a genetic switch toggling from OFF to ON [7], where hundreds of thousands of simulations can be needed to get reasonable relative error bounds.

The evolution of the probability distribution on the state space over time is precisely described by the *chemical master equation* (CME) [8]; however, it is computationally difficult to solve the CME precisely for systems with large state spaces. The need to precisely describe the statistics of rare events motivated the development of methods for finding approximate solutions to the CME, such as the finite state projection method [7] or uniformization [9].

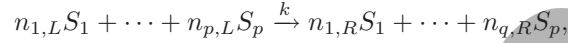
Methods for exactly or approximately calculating the solution to a CME determine the *a priori* probability distributions of events occurring in the process at particular instants in time. However, there is also an *a posteriori* problem that we consider in this paper: Based on observed data from a chemical process and a continuous-time Markov process model of its behavior, what is the probability of a rare event having occurred? This problem is non-trivial because our observation of biochemical processes is very limited. In practice, it is possible to estimate the populations of proteins and other biochemical species *in vivo* by tagging the species of interest with fluorescent markers. Even so, in order to detect the populations of multiple species we must tag them so that they fluoresce at different wavelengths, thus it is difficult to observe the populations of more than four species at any one time without the emission spectra overlapping each other.

Single-cell studies allow the researcher in the laboratory to collect stochastic, dynamic data from the behavior of an intracellular reaction network [2]. However, the limited observation capabilities available to in the lab motivate the need for intelligent algorithms for analyzing this data to determine the probabilities that events of interests have occurred. In this paper, we develop an approach for using the CME description of a chemical process in order to determine the *a posteriori* probabilities that a particular event of interest has occurred. The methodology we use is based on techniques for addressing the problem of diagnosability in discrete event system (DES) models [10]; in particular, stochastic DES models [11]. The techniques of DES diagnosis allow us to develop a method for calculating the probabilities of rare events based on dynamic data; thus, they are generally applicable to systems such as toggle switches and oscillators in which a steady-state behavior is never realized.

We organize the paper as follows. In Section 2 we define the continuous-time Markov process model and state the diagnosis problem. In Section 3 we derive a hybrid system representation of the *a posteriori* probability distribution for the case when our observations of the system are intermittent; in section 4, we do the same for the case of continuous observations. In Section 5, we demonstrate how calculating this *a posteriori* distribution can be used to solve the diagnosis problem. We demonstrate the performance of the algorithm on a stochastic gene expression model in Section 6.

2 Problem Formulation

Consider a reaction chamber containing a set of species $\mathcal{S} = \{S_1, S_2, \dots, S_n\}$ that interact along a set of reaction channels \mathcal{R} . Each reaction in \mathcal{R} takes the form



where $S_i \in \mathcal{S}$, $n_{i,L}, n_{i,R} \in \mathbb{Z}^{\geq 0}$, $i = 1 \dots p$. The constant k above the arrow is the base rate of the reaction. In the paper, we make the standard assumptions of stochastic chemical kinetics [12], namely that

- the reaction chamber is of fixed volume,
- the reaction chamber is at thermal equilibrium,
- the particles in the reaction chamber are well-mixed, i.e., the rate at which non-reacting collisions between particles occur is much greater than the rate at which reacting collisions occur.

Under these standard conditions, the state of the reaction network can be described at the mesoscopic scale as a p -dimensional vector $x(t) = [N_1(t) \dots N_p(t)]^T$, where $N_i(t)$ denotes the number of the species S_i at time t . Furthermore, under these conditions it has been rigorously derived that the evolution of the system state can be precisely described using a continuous time Markov process (CTMP) [6].

Definition 1. A continuous time Markov process is a tuple $\mathcal{S} = (X, \mathbf{Q}, \pi_0)$, where X is a countable set of states, \mathbf{Q} is a transition rate matrix, and π_0 is the initial probability distribution on X .

The elements of \mathbf{Q} specify the rates at which transitions in the CTMP occur. If \mathbf{Q}_{ij} is a non-diagonal element of \mathbf{Q} , then the probability of a transition from a state $x_i \in X$ to another state $x_j \in X$ in the interval $[t, t + dt)$ is $\mathbf{Q}_{ij}dt$; if \mathbf{Q}_{ii} is a diagonal element of \mathbf{Q} , we set that $\mathbf{Q}_{ii} = -\sum_{j \neq i} \mathbf{Q}_{ij}$, thereby ensuring that all of the columns in \mathbf{Q} sum to zero.

A CTMP produces trajectories that are piecewise constant, right-continuous functions $\omega : \mathcal{T} \rightarrow X$, where $\mathcal{T} \triangleq [0, t_{max}]$. We say that a state x is visited along the trajectory ω if there exists $t \leq t_{max}$ such that $\omega(t) = x$.

Chemical Master Equation. Enumerate the elements of the state space X as $X = \{x_1, x_2, \dots, x_n, \dots\}$. Denote by $\mathbf{p}(t)$ a vector where the i^{th} element is the probability that the system is in state x_i at time t , i.e. $\mathbf{p}_i(t) \triangleq \Pr(\omega(t) = x_i)$. The *chemical master equation* (CME) describes how the probability vector $\mathbf{p}(t)$ evolves with time. The CME is expressed in vector form as a linear system of equations [8]:

$$\dot{\mathbf{p}}(t) = \mathbf{Q}\mathbf{p}(t). \quad (1)$$

Because the initial condition of this ODE is $\mathbf{p}(0) = \pi_0$, the solution of the CME is $\mathbf{p}(t) = e^{\mathbf{Q}t}\pi_0$.

Intermittent Observation Model. In practice, it is not possible to completely observe the state of the CTMP \mathcal{S} ; only some of the state variables are available for observation as there do not exist methods to precisely measure

the numbers of many types of biochemical species. To model this partial observation of the state, we define a set of outputs Y and we define a *state output random variable* $h : X \rightarrow Y$. For each $y \in Y$, we define a vector \mathbf{h}_y by $(\mathbf{h}_y)_i \triangleq \Pr(h(x_i) = y)$. It will be convenient to express the output probabilities in a matrix form, so we also define $\mathbf{H}_y \triangleq \text{diag}(\mathbf{h}_y)$.

As the CTMP \mathcal{S} evolves with time, we make n observations at the sequence of times $T_n \triangleq \{\tau_1, \tau_2, \dots, \tau_n\} \subset \mathcal{T}$. The results in this paper hold when the sample times in T_n are either periodic or aperiodic. Denote by y_k the output value observed at time τ_k . Denote by y^k the sequence of observed values $\{y_1, y_2, \dots, y_k\}$.

Continuous Observation Model. We also consider the case that the behavior of the CTMP \mathcal{S} is observed continuously, that is, the value of $h(\omega(t))$ is known for all $t \in \mathcal{T}$. In the continuous observation case, we make the assumption that the state output is deterministic, that is, for all $x \in X$ there exists an $y \in Y$ such that $\Pr(h(x) = y) = 1$.

We denote by $y^{[0,t]}$ the set of observed values made on the interval $[0, t]$. Because trajectories of the CTMP are piecewise constant functions, the output $y^{[0,t]}$ is also piecewise constant. If $y^{[0,t]}$ consists of n distinct pieces, we denote by $\{y_1, y_2, \dots, y_n\}$ the values of the n distinctly observed outputs and we denote by $\{\tau_1, \tau_2, \dots, \tau_n\}$ the times at which the n jumps from one output to another occurred.

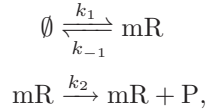
For each output y_i , we define by $I = \{i_1, i_2, \dots, i_n\}$ the index set of states in X whose output is y_i with probability 1. For two outputs y_i and y_j , we denote by \mathbf{Q}_{IJ} the submatrix of the transition rate matrix where the rows are selected according to the index set I and the columns are selected according to the index set J .

Diagnosis Problems: The problem we consider in this paper is the following. Let $X_S \subset X$ denote a set of *special states*. In practice, the event that the system enters a state in X_S could correspond to several different occurrences of biological interest. For example, the set X_S can be defined so that the event corresponds to the population of a particular species in \mathcal{S} crossing a given threshold, thereby activating a downstream genetic circuit or flipping the state of an oscillator or toggle switch.

We denote by D_t the event that a trajectory of the CTMP \mathcal{S} has reached a special state before a given time t , and formally define $D_t \triangleq \{\omega : \exists s \leq t \text{ such that } \omega(s) \in X_S\}$ for all $t \in \mathcal{T}$.

We consider two diagnosis problems. The *intermittent observation diagnosis problem* is: given a sequence of noisy observations y^n made at time $T_n = \{\tau_1, \tau_2, \dots, \tau_n\}$, find $\Pr(D_t | y^n)$ for all $t \geq \tau_n$. Similarly, the *continuous observation diagnosis problem* is: for all $t > 0$, given the perfect observations $y^{[0,t]}$ made on the interval $[0, t]$, find $\Pr(D_t | y^{[0,t]})$.

Example 1. Consider the following reaction network model of stochastic gene expression:



where the symbol mR denotes messenger RNA and the symbol P denotes fluorescent protein. The rate constants are selected to be $k_1 = .0554$ mRNA/s, $k_2 = .17$ protein/(mRNA.s) and $k_{-1} = .0113$ (mRNA.s)⁻¹ [13]. For simplicity of presentation, we assume that rate of protein decay is on a slower time scale than the other reactions and thus omit this reaction; the reaction can be added to the model in a straightforward manner. Interpreting this reaction network stochastically produces a CTMP with $X = \mathbb{Z}^{\geq 0} \times \mathbb{Z}^{\geq 0}$, where a typical state is of the form $x = [n_{mR} \ n_P]$. We are interested in investigating the behavior of this system until $t = 1440$ seconds, the approximate time at which cell division occurs [4].

The state space X we defined is infinite. For the sake of calculation, we assume that the probabilities that the populations of mRNA and protein ever become very large are negligible. We thus define a constant mR_{max} and specify that the probability that another mRNA is created when the current mRNA population is mR_{max} is zero. We also define a constant P_{max} that plays the same role for protein. The state space we define in this manner is $X = \{0, 1, \dots, mR_{max}\} \times \{0, 1, \dots, P_{max}\}$.

To construct the transition rate matrix \mathbf{Q} , we define an indexing function $\mathcal{I} : X \rightarrow \mathbb{N}$. The non-zero elements of \mathbf{Q} are specified to be

$$\begin{aligned} \mathbf{Q}_{\mathcal{I}(n_{mR}+1, n_P), \mathcal{I}(n_{mR}, n_P)} &= k_1 && \text{if } n_{mR} < mR_{max} \\ \mathbf{Q}_{\mathcal{I}(n_{mR}-1, n_P), \mathcal{I}(n_{mR}, n_P)} &= k_{-1} n_{mR} && \text{if } n_{mR} > 0 \\ \mathbf{Q}_{\mathcal{I}(n_{mR}, n_P+1), \mathcal{I}(n_{mR}, n_P)} &= k_2 n_{mR} && \text{if } n_P < P_{max}. \end{aligned}$$

All other elements of \mathbf{Q} are zero.

The initial distribution π_0 is defined by observing that the protein number at $t = 0$ is zero. We assume the mRNA number is distributed according to a Poisson distribution with parameter $\lambda = 5$.

We consider two output functions. The noiseless output function is given by $\Pr(h([n_{mR} \ n_P]) = n_P) = 1$, i.e. the fluorescent protein number is observed perfectly. The noisy output function is $\Pr(h([n_{mR} \ n_P]) = n_P - \ell) = \frac{1}{11}$ for $\ell \in \{-5, -4, \dots, 4, 5\}$. That is, under noisy observations the observed protein number is uniformly distributed around the true protein number in an interval of ± 5 proteins. This noise model is used for the purposes of illustration and other noise models, such as Gaussian or Poisson noise, can be substituted.

We define the set of special states as $X_S = \{[n_{mR} \ n_P] : n_{mR} \geq 9\}$. The diagnosis problems are then to find the probability that at some point on the interval $[0, t]$, the population of mRNA was at least 9 given a sequence of observations (y^n for the intermittent observation case, $y^{[0,t]}$ for the continuous observation case.)

3 The *A Posteriori* Probability Distribution Under Intermittent Observations

In order to solve the intermittent observation diagnosis problem, we first propose a more general question. Given a sequence of observations y^n made at times $T_n = \{\tau_1, \dots, \tau_n\}$, what is the probability of being in a given state at time $t \geq \tau_n$? Expressed more compactly, the problem is to find the *a posteriori* probability distribution $\mathbf{p}(t | y^n)$.

Theorem 1. *For $t \geq \tau_n$, the a posteriori probability distribution $\mathbf{p}(t | y^n)$ is*

$$\mathbf{p}(t | y^n) = \frac{1}{K} e^{\mathbf{Q}(t-\tau_n)} \mathbf{H}_{y_n} e^{\mathbf{Q}(\tau_n-\tau_{n-1})} \mathbf{H}_{y_{n-1}} \dots \mathbf{H}_{y_2} e^{\mathbf{Q}(\tau_2-\tau_1)} \mathbf{H}_{y_1} e^{\mathbf{Q}(\tau_1)} \pi_0. \quad (2)$$

Proof. We find $\mathbf{p}(t | y^n)$ by decomposing the evolution of the *a posteriori* probability distribution into two domains. Between observation times, we show that $\mathbf{p}(t | y^n)$ evolves continuously according to the equation

$$\dot{\mathbf{p}}(t | y^n) = \mathbf{Q} \mathbf{p}(t | y^n). \quad (3)$$

At the times T_n when observation occurs, we show that $\mathbf{p}(t | y^n)$ evolves according to a discrete jump

$$\mathbf{p}(\tau_n^+ | y^n) = \frac{1}{K_n} \mathbf{H}_{y_n} \mathbf{p}(\tau_n^- | y^{n-1}). \quad (4)$$

For brevity, we denote the event $\{\omega(t) = x\}$ by $[t, x]$. To describe the evolution of the state *between observations*, we condition on the value of the state at τ_n , the time of the most recent observation, yielding the equation

$$\begin{aligned} \Pr([t, x_j] | y^n) &= \sum_{x_i \in X} \Pr([t, x_j], [\tau_n, x_i] | y^n) \\ &= \sum_{x_i \in X} \Pr([t, x_j] | [\tau_n, x_i], y^n) \Pr([\tau_n, x_i] | y^n). \end{aligned}$$

The right hand side of the above equation is the expression for calculating one element in a matrix multiplication. Constructing the vector $\mathbf{p}(t | y^n)$ from this expression yields

$$\begin{aligned} \mathbf{p}(t | y^n) &= \begin{bmatrix} \Pr([t, x_1] | [\tau_n, x_1]) \dots \Pr([t, x_1] | [\tau_n, x_k]) \\ \vdots \quad \ddots \quad \vdots \\ \Pr([t, x_k] | [\tau_n, x_1]) \dots \Pr([t, x_k] | [\tau_n, x_k]) \end{bmatrix} \begin{bmatrix} \Pr([\tau_n, x_1] | y^n) \\ \vdots \\ \Pr([\tau_n, x_k] | y^n) \end{bmatrix}, \\ &= e^{\mathbf{Q}(t-\tau_n)} \mathbf{p}(\tau_n | y^n), \end{aligned} \quad (5)$$

where the last line follows from the solution to the CME.

To calculate how the probability distribution updates discretely *at observation times*, we condition of the state of the process at a time $\tau_n - dt$ for small dt .

$$\begin{aligned} \Pr([\tau_n^+, x_j] | y^n) &= \sum_{x_i \in X} \Pr([\tau_n^+, x_j], [\tau_n - dt, x_i] | y^n) \\ \Pr([\tau_n^+, x_j] | y^n) &= \frac{1}{\Pr(y_n | y^{n-1})} \sum_{x_i \in X} \Pr([\tau_n^+, x_j], [\tau_n - dt, x_i], y_n | y^{n-1}) \\ \Pr([\tau_n^+, x_j] | y^n) &= \frac{1}{K_n} \sum_{x_i \in X} \Pr([\tau_n^+, x_j], y_n | [\tau_n - dt, x_i]) \Pr([\tau_n - dt, x_i] | y^{n-1}), \end{aligned} \quad (6)$$

where $K_n = \Pr(y_n | y^{n-1})$. For small dt , the quantity $\Pr([\tau_n^+, x_j], y_n | [\tau_n - dt, x_i])$ reduces to $\mathbf{Q}_{ij} dt \Pr(h(x_j) = y_n)$ if $i \neq j$; if $i = j$, the term reduces to $(1 - \mathbf{Q}_{ii} dt) \Pr(h(x_i) = y_n)$. As $dt \rightarrow 0$, the first of these expressions goes to zero while the second goes to $\Pr(h(x_i) = y_n)$, which is equal to $(\mathbf{h}_{y_n})_i$. Therefore

$$\Pr([\tau_n^+, x_j] | y^n) = \begin{cases} \frac{1}{K_n} (\mathbf{h}_{y_n})_i \Pr([\tau_n^-, x_i] | y^{n-1}) & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

or in vector form,

$$\mathbf{p}(\tau_n^+ | y^n) = \frac{1}{K_n} \mathbf{H}_{y_n} \mathbf{p}(\tau_n^- | y^{n-1}).$$

Because $\mathbf{p}(\tau_n^+ | y^n)$ is a probability distribution, it follows that K_n is a normalization constant and thus, $K_n = \mathbf{1}^T \mathbf{H}_{y_n} \mathbf{p}(\tau_n^- | y^{n-1})$.

Combining the results of the discrete and continuous cases yields

$$\mathbf{p}(t | y^n) = \frac{1}{K_n} e^{\mathbf{Q}(t-\tau_n)} \mathbf{H}_{y_n} \mathbf{p}(\tau_n^- | y^{n-1}).$$

Similarly we can show that $\mathbf{p}(\tau_k^- | y^{k-1}) = \frac{1}{K_{k-1}} e^{\mathbf{Q}(\tau_k - \tau_{k-1})} \mathbf{H}_{y_{k-1}} \mathbf{p}(\tau_{k-1}^- | y^{k-2})$ for all $2 \leq k \leq n-1$. For $k=1$, $\mathbf{p}(\tau_1^- | y^0) = e^{\mathbf{Q}\tau_1} \pi_0$ as a result of the *a priori* CME; because there are no observations before τ_1 , the conditional and unconditional distributions are equal. Combining all these results yields the desired result

$$\mathbf{p}(t | y^n) = \frac{1}{K} e^{\mathbf{Q}(t-\tau_n)} \mathbf{H}_{y_n} e^{\mathbf{Q}(\tau_n - \tau_{n-1})} \mathbf{H}_{y_{n-1}} \dots \mathbf{H}_{y_2} e^{\mathbf{Q}(\tau_2 - \tau_1)} \mathbf{H}_{y_1} e^{\mathbf{Q}\tau_1} \pi_0,$$

where $K = K_1 K_2 \dots K_n$. □

Because the *a posteriori* probability distribution evolves continuous except at the n observation times, where it jumps discretely, it is natural to describe its evolution using a hybrid system.

Theorem 2. *Given a sequence of observations y_n made at times T_n , the a posteriori probability distribution is described by the following dynamics:*

$$\dot{\mathbf{p}} = \mathbf{Q}\mathbf{p} \quad (\text{continuous dynamics}) \quad (7)$$

$$\mathbf{p} \mapsto \frac{\mathbf{H}_{y_n}\mathbf{p}}{\mathbf{1}^T\mathbf{H}_{y_n}\mathbf{p}}. \quad (\text{discrete dynamics}) \quad (8)$$

The system evolves according to the continuous dynamics in the set \mathcal{T}/T_n and according to the discrete dynamics in the set $D = T_n$. The initial distribution of $\mathbf{p}(t | y^n)$ is π_0 .

The correctness of this representation can be proven by simple verification of the solution in Equation 2.

The hybrid system representation reveals that between observations, the system evolves according to the standard CME (Equation 1). When an observation is made, the probability distribution on the states is reweighted so that states that are more likely to have produced the observation value increase in probability while those that are less likely decrease in probability.

Except for the normalization factor $\mathbf{1}^T\mathbf{H}_{y_n}\mathbf{p}$, the hybrid system (7)-(8) is linear. We can choose to disregard the normalization factor because the unnormalized probability vector contains the same information about the distribution of the states as the normalized vector and thus make the system wholly linear. Alternatively, we can treat the system as a Weiner model, as it is linear except for the static normalization linearity that appears just before the output.

4 The *A Posteriori* Probability Distribution Under Continuous Observations

The situation is analogous when the objective is to calculate the *a posteriori* probability distribution $\mathbf{p}(t | y^{[0,t]})$ under continuous observations.

Theorem 3. *For $t \geq \tau_n$, the a posteriori probability distribution $\mathbf{p}(t | y^{[0,t]})$ given a sequence of observations y_1, y_2, \dots, y_n and a sequence of jump times $T_n = \{\tau_1, \tau_2, \dots, \tau_n\}$ is*

$$\mathbf{p}(t | y^{[0,t]}) = \frac{1}{K} e^{\mathbf{Q}_{I_n I_n}(t-\tau_n)} \mathbf{Q}_{I_n I_{n-1}} e^{\mathbf{Q}_{I_{n-1} I_{n-1}}(\tau_n-\tau_{n-1})} \mathbf{Q}_{I_{n-1} I_{n-2}} \dots \dots \mathbf{Q}_{I_3 I_2} e^{\mathbf{Q}_{I_2 I_2}(\tau_2-\tau_1)} \mathbf{Q}_{I_2 I_1} e^{\mathbf{Q}_{I_1 I_1}(\tau_1)} \pi_0. \quad (9)$$

We omit the full proof of this theorem for reasons of space; instead, we sketch the key differences between the proofs of Theorems 1 and 3. We proceed by showing that between jump times, the distribution $\mathbf{p}(t | y^{[0,t]})$ evolves continuously according to the equation

$$\mathbf{p}(t | y^{[0,t]}) = \frac{1}{K_{nc}} e^{\mathbf{Q}_{II}(t-\tau_n)} \mathbf{p}(\tau_n | y^{[0,\tau_n]}), \quad (10)$$

and that at jump times, $\mathbf{p}(t | \mathbf{y}^{[0,t]})$ evolves according to the discrete jump

$$\mathbf{p}(\tau_n^+ | \mathbf{y}^{[0,\tau_n]}) = \frac{1}{K_{nd}} \mathbf{Q}_{JIP}(\tau_n^- | \mathbf{y}^{[0,\tau_n]}). \quad (11)$$

The proof in the continuous domain follows analogous steps until Equation 5 is reached. The general term in the $n \times n$ matrix in the continuous case is $\Pr([t, x_1], \mathbf{y}^{[\tau_n, t]} \equiv y_i | [\tau_n, x_1])$; the condition $\{\mathbf{y}^{[\tau_n, t]} \equiv y_i\}$ is now added as we know the output value at all times between the jumps. To find these probabilities we do not solve the full CME $\dot{\mathbf{p}} = \mathbf{Q}\mathbf{p}$; instead, we solve the reduced CME $\dot{\mathbf{p}} = \mathbf{Q}_{II}\mathbf{p}$ to give us the probability of the set of trajectories that have the output y_i at every instant along the interval $[\tau_n, t]$. Because the columns of \mathbf{Q}_{II} can sum to less than zero, we also require the introduction of the normalization constant $K_{nc} = \Pr(\mathbf{y}^{[\tau_n, t]} \equiv y_i | \mathbf{y}^{[0,\tau_n]})$.

The proof in the discrete domain is identical until Equation 6 is reached. The normalization constant K_{nd} is given by $K_{nd} = \Pr(\mathbf{y}_{\tau_n} | \mathbf{y}^{[0,\tau_n-dt]})$, which is of the order of dt . The probability $\Pr([\tau_n^+, x_j], \mathbf{y}_{\{\tau_n\}} | [\tau_n - dt, x_i])$ on the right hand side is necessarily $\mathbf{Q}_{ji}dt$ because we know the output is y_i before the jump and y_j after the jump. We can cancel out the dt with the dt in K_{nd} , yielding the result.

The continuous observation *a posteriori* probability distribution is also expressible as a hybrid system.

Theorem 4. *The a posteriori probability distribution is described by the following dynamics:*

$$\dot{\mathbf{p}} = \mathbf{Q}_{II}\mathbf{p} - (\mathbf{1}^T \mathbf{Q}_{II}\mathbf{p}) \mathbf{p} \quad (\text{continuous dynamics}) \quad (12)$$

$$\mathbf{p} \mapsto \frac{\mathbf{Q}_{JIP}}{\mathbf{1}^T \mathbf{Q}_{JIP}}. \quad (\text{discrete dynamics}) \quad (13)$$

As in the intermittent observation case, the system evolves according to the continuous dynamics in the set \mathcal{T}/T_n and according to the discrete dynamics in the set $D = T_n$. The initial distribution of $\mathbf{p}(t | \mathbf{y}^n)$ is π_0 .

Whereas in the intermittent observation case, normalization was only necessary in the discrete domain T_n , in the continuous observation case normalization is necessary in both domains. The term $(\mathbf{1}^T \mathbf{Q}_{II}\mathbf{p}) \mathbf{p}$ corrects for the need for constant normalization.

5 Solution to the Diagnosis Problem

Having developed equations for describing the evolution of the *a posteriori* probability distributions under both continuous and intermittent observations, we now return our attention to the diagnosis problem. Recall that D_t denotes the event that a state $x \in X_S$ was visited along the interval $[0, t]$; our objective is to find the probabilities $\Pr(D_t | \mathbf{y}^n)$ and $\Pr(D_t | \mathbf{y}^{[0,t]})$. To find these probabilities, we construct an extended continuous time Markov process from our original process \mathcal{S} and calculate the *a posteriori* probability distributions for this extended process.

Let S be an index set marking the indices of the special states X_S , and denote by N the complement of S . After partitioning X in this manner, we write \mathbf{Q} as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{NN} & \mathbf{Q}_{SN} \\ \mathbf{Q}_{NS} & \mathbf{Q}_{SS} \end{bmatrix},$$

where \mathbf{Q}_{NN} defines the transition rates within the normal states, \mathbf{Q}_{SS} defines the transition rates within the special states, and \mathbf{Q}_{NS} and \mathbf{Q}_{SN} define the transitions rates between the normal and special states.

For any CTMP \mathcal{S} equipped with a set special states X_S , we define an extended CTMP $\bar{\mathcal{S}} = (X \cup X_{N'}, \bar{\mathbf{Q}}, \bar{\pi}_0)$. We construct the state space of $\bar{\mathcal{S}}$ by appending to the state space of \mathcal{S} the set of extended states $X_{N'}$. Each state in $X_{N'}$ corresponds to a normal state that is visited by a trajectory *after* a special state has been visited. This is seen in extended transition rate matrix, which is defined to be

$$\bar{\mathbf{Q}} \triangleq \begin{bmatrix} \mathbf{Q}_{NN} & 0 & 0 \\ \mathbf{Q}_{NS} & \mathbf{Q}_{SS} & \mathbf{Q}_{NS} \\ 0 & \mathbf{Q}_{SN} & \mathbf{Q}_{NN} \end{bmatrix}.$$

The extended initial distribution is defined as $\bar{\pi}_0(x) = \pi_0(x)$ if $x \in X$ and $\bar{\pi}_0(x) = 0$ if $x \in X_{N'}$. Similarly, we define the observation model of the extended states. For all $y \in Y$, $\Pr(\bar{h}(x) = y) \triangleq \Pr(h(x) = y)$ if $x \in X$; for an extended state $x_{N'_i}$, $\Pr(\bar{h}(x_{N'_i}) = y) \triangleq \Pr(h(x_{N_i}) = y)$, i.e., each extended state has the same output distribution as its corresponding normal state.

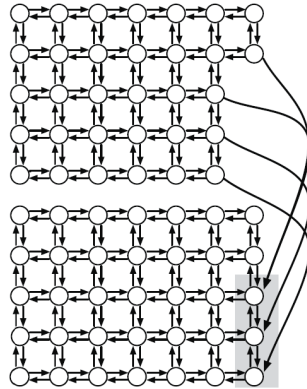


Fig. 1. Construction of the extended state space. Transitions from the special states (shaded) no longer reach the normal states (top) and are instead directed to the extended states (bottom, unshaded). The probability of having visited a special state is thus the probability of currently being in either a special state or extended state. Figure modified from [14]

By applying the *a posteriori* solution to the CME to the extended CTMP \bar{S} , we solve the diagnosis problem.

Theorem 5. *The a posteriori probability of D_t given a sequence of observations at times $T_n = \{\tau_1, \tau_2, \dots, \tau_n\}$, is*

$$\Pr(D_t | y^n) = \mathbf{1}^T \mathbf{p}_S(t | y^n) + \mathbf{1}^T \mathbf{p}_{N'}(t | y^n). \quad (14)$$

Proof. Clearly $\omega(t) \in X_S$ implies D_t . By construction of \bar{S} , if $\omega(t) \in X_{N'}$, then there exists $s < t$ where $\omega(s) \in X_S$ because $\bar{\pi}_0(x) = 0$ for all $x \in X_{N'}$ and there does not exist a path from X_N to $X_{N'}$ that does not pass through X_S . Therefore $\omega(t) \in X_{N'}$ implies D_t . However, if $\omega(t) \in X_N$, there cannot exist such an s because there is no path from X_S to X_N . Therefore the event D_t is equal to the event $\omega(t) \in (X_S \cup X_{N'})$, whose probability is given by the right hand side of Equation 14. \square

The probabilities on the right hand side of Equation 14 can be calculated using Equations 7-8. The analogous result holds for the case of continuous observations.

Theorem 6. *The a posteriori probability of D_t given the observations $y^{[0,t]}$ is*

$$\Pr(D_t | y^{[0,t]}) = \mathbf{1}^T \mathbf{p}_S(t | y^{[0,t]}) + \mathbf{1}^T \mathbf{p}_{N'}(t | y^{[0,t]}). \quad (15)$$

Similarly, the probabilities on the right hand side of Equation 15 can be calculated using Equations 12-13.

6 Diagnosing mRNA Levels in Stochastic Gene Expression

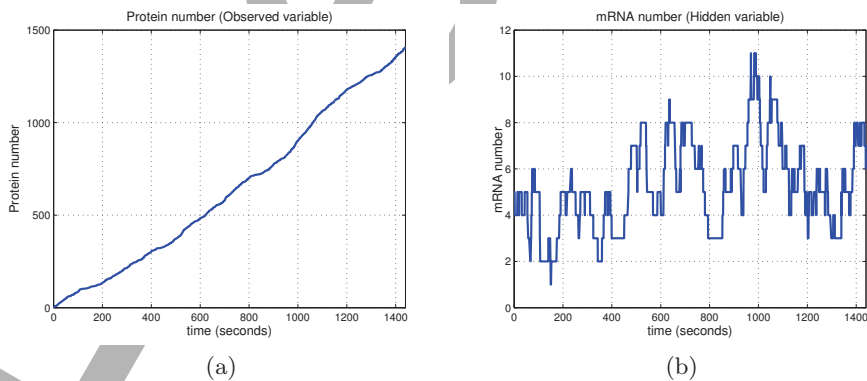


Fig. 2. A simulated trajectory of the stochastic gene expression reaction network. (a) The observed output of the reaction network is the fluorescent protein number. (b) The unobserved state variable is the mRNA number.

We now return to the stochastic gene expression model defined in Example 1. Using Gillespie's stochastic simulation algorithm [5], we generate a trajectory $\omega(t)$ along the interval $\mathcal{T} = [0, 1440]$. The trajectory is shown in Figure 2.

We first consider the case of continuous observation. The protein number, shown in Figure 2(a) is observed for all times. The mRNA number, shown in Figure 2(b) is unobserved but is shown for comparison. By inspecting the evolution of the mRNA number, we can conclude that the set of special states is first reached when the mRNA number is first equal to 9, which occurs at approximately $t = 620$ seconds.

Using the observed protein number as the input, we evaluate the diagnoser dynamics under continuous observation using Equations 12 - 13. The evolution of the diagnoser output over time is shown in Figure 3. When no jump is observed

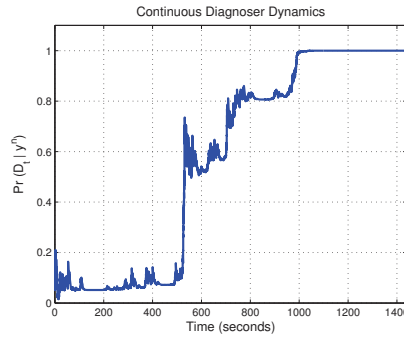


Fig. 3. The continuous observation diagnoser dynamics.

in the protein number, the probability that the system was ever in a special state decreases because states in X_S have high mRNA numbers. Thus it is less likely that the protein level remains constant in a special state, and the *a posteriori* probability of having been in X_S decreases. Similarly, when a jump is observed, the probability of having been in a special state increases, because jumps are more likely to have occurred from special states than from normal ones. Notice that the largest jumps in the $\mathbf{p}(D_t | y^{[0,t]})$ correspond to the fastest increases in the protein number. The first large increase in $\mathbf{p}(D_t | y^{[0,t]})$ actually occurs before the special states are first reached; by inspecting the trajectory, we can see that the system is at the boundary of the normal states (the mRNA number is 8) when the increase occurs and that the rate at which protein is being produced is very high.

In the case of intermittent observations, we consider three different sequences of sampling times corresponding to fixed sampling intervals of 10, 60, and 120 seconds, respectively. The diagnoser dynamics are shown in Figure 4(a). In the intermittent observation case, the $\mathbf{p}(D_t | y^n)$ always increases between observations because probability mass is flowing from the normal states to the special

states but there is no path by which it can return. When observations occur, the probability of D_t increases if the change in protein number since the last observation is large and decreases if it is small. As the sampling interval decreases, the diagnoser dynamics approach the dynamics seen in the continuous observation case.

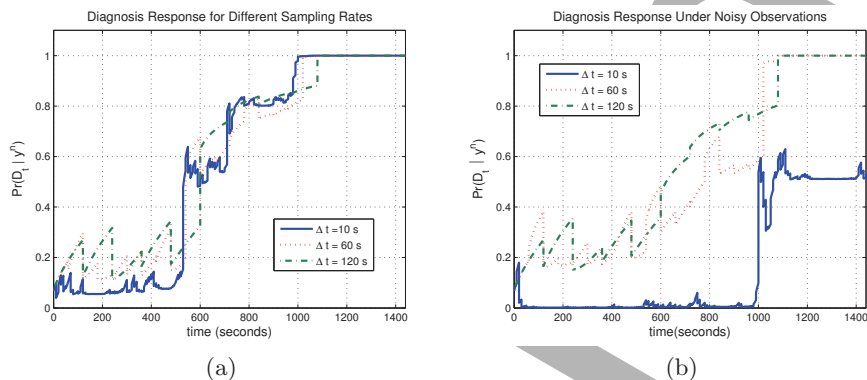


Fig. 4. The intermittent observation diagnoser dynamics for three different sampling periods. (a) The protein number is observed without noise. (b) The protein number is observed with noise.

The same result does not occur when the observations are noisy, as shown in Figure 4(b). In the diagnoser dynamics for the sampling interval of 10 seconds, the probability $\mathbf{p}(D_t | y^n)$ does not increase when the first mRNA number first goes high at $t = 620$ seconds. The diagnoser responds when the mRNA number goes high for a second time at 1000 seconds at the final probability of D_t is approximately 53%. Interestingly, the dynamics with slower sampling intervals perform much more similarly to the noiseless case. Prefiltering the raw observations before evaluating the *a posteriori* probabilities is likely to improve the performance of the diagnosis response when confronted with noisy data.

7 Discussion

In this paper, we consider the problem of finding the *a posteriori* probability that a rare event of interest has occurred in a biochemical process, given a record of observations of that process. We derive equations for determining this probability for the case where the observations are made continuously and the case where the observations are made intermittently.

We illustrated the approach using simulated data of a model of stochastic gene expression. We are currently investigating the applicability of the approach to real data collected from single-cell microscopy studies of a synthetic genetic circuit in *E. coli*. We expect the experimental findings will motivate the need for

further theoretical research such as studying the effect of model perturbations on the diagnosis performance and developing strict guarantees as to the magnitude of the error introduced by eliminating unlikely states from the CTMP model.

This research is supported by the 2006 AFOSR MURI award “High Confidence Design for Distributed Embedded Systems” and an A. Richard Newton Breakthrough Research Award from Microsoft. The author wishes to acknowledge E. Klavins and D. Georgiev, University of Washington, for fruitful discussions and constructive criticism and B. Munsy, Los Alamos National Laboratory, for permission to use Figure 1.

References

1. Mettetal, J., van Oudenaarden, A.: Necessary noise. *Science* **317**(5837) (July 2007) 463–464
2. Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S.: Stochastic Gene Expression in a Single Cell. *Science* **297**(5584) (2002) 1183–1186
3. McAdams, H., Arkin, A.: Stochastic mechanisms in gene expressions. *Proc. Natl. Acad. Sci.* **94** (February 1997) 814–819
4. Swain, P., Elowitz, M., Siggia, E.: Intrinsic and extrinsic contributions to stochasticity in gene expression. *Science* **99**(20) (2002) 12795–12800
5. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. *J. of Phys. Chem.* **81** (1977) 2340–2360
6. Gillespie, D.T.: Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry* **58**(1) (2007) 35–55
7. Munsy, B., Khammash, M.: The finite state projection algorithm for the solution of the chemical master equation. *J. Chemical Physics* **124**(044104) (2006)
8. Van Kampen, N.G.: *Stochastic Processes in Physics and Chemistry*, Third Edition (North-Holland Personal Library). North Holland (2007)
9. Sandmann, W., Wolf, V.: Computational probability for systems biology. In: *Lecture Notes in Computer Science*, vol. 5054, *Proceedings of Formal Methods in Systems Biology*, Springer-Verlag Inc. (2008) 33–47
10. Lafortune, S., Teneketzis, D., Sampath, M., Sengupta, R., K.Sinnamohideen: Failure diagnosis of dynamic systems: An approach based on discrete event systems. In: *Proc. 2001 American Control Conference*. (June 2001) 2058–2071
11. Thorsley, D., Teneketzis, D.: Diagnosability of stochastic discrete-event systems. *IEEE Trans. Automatic Control* **50**(4) (April 2005) 476–492
12. McQuarrie, D.: Stochastic approach to chemical kinetics. *J. Applied Probability* **4** (1967) 413–478
13. Thorsley, D., Klavins, E.: Model reduction of stochastic processes using Wasserstein pseudometrics. In: *Proceedings of the 2008 American Control Conference*. (2008) 1374–1381
14. Munsy, B., Khammash, M.: Computation of switch time distributions in stochastic gene regulatory networks. In: *Proc. 2008 American Control Conference*. (June 2008) 2761–2766