Population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression

Adam R. Stinchcombe, Charles S. Peskin, and Daniel Tranchina

Phys. Rev. E **85**, 061919 — Published 22 June 2012

DOI: 10.1103/PhysRevE.85.061919

A population density approach for discrete mRNA distributions in generalized switching models for stochastic gene expression

Adam R. Stinchcombe,* Charles S. Peskin,† and Daniel Tranchina‡
Courant Institute of Mathematical Sciences, New York University,
251 Mercer Street, New York, New York 10012, USA
(Dated: June 7, 2012)

We present a generalization of a population density approach for modeling and analysis of stochastic gene expression. In the model, the gene of interest fluctuates stochastically between an inactive state, in which transcription cannot occur, and an active state, in which discrete transcription events occur; and the individual mRNA molecules are degraded stochastically in an independent manner. This sort of model in simplest form with exponential dwell times has been used to explain experimental estimates of the discrete distribution of random mRNA copy number. In our generalization, the random dwell times in the inactive and active states, T_0 and T_1 respectively, are independent random variables drawn from any specified distributions. Consequently, the probability per unit time of switching out of a state depends on the time since entering that state. Our method exploits a connection between the fully discrete random process and a related continuous process. We present numerical methods for computing steady-state mRNA distributions and an analytical derivation of the mRNA autocovariance function. We find that empirical estimates of the steady-state mRNA probability mass function from Monte Carlo simulations of labaratory data do not allow one to distinguish between underlying models with exponential and non-exponential dwell times in some relevant parameter regimes. However, in these parameter regimes and where the autocovariance function has negative lobes, the autocovariance function disambiguates the two types of models. Our results strongly suggest that temporal data beyond the autocovariance function is required in general to characterize gene switching.

PACS numbers: 87.10.Mn

* stinch@courant.nyu.edu
† peskin@courant.nyu.edu
‡ tranchina@courant.nyu.edu; Department of Biology, New York University, New York, New York 10003, USA

I. INTRODUCTION

There is a recent surge of interest in stochastic gene expression and its implications for random cell fate in development (reviewed by Johnston and Desplan [1]) and stochastic switching of cell phenotype [2–10].

We present here a generalized model for stochastic burst-like transcription. This model in its simplest form was used by Raj *et al.* [11] to explain and quantify their experimental estimate of the distribution of random mRNA copy number, for several different transcripts, over a population of genetically homogeneous eukaryotic cells in culture. We will take a population density approach in which a direct calculation of a probability density function describes the distribution of state variables across a population (ensemble) of cells.

Variability among individuals in a population of organisms arises from a number of causes. Those most commonly thought of are genetic variation, in which differences in DNA sequence lead to different phenotypes, and environmental variation, in which individual organisms alter their behavior in different surroundings and conditions. Recent studies speak to the question of what would happen if those sources of variability were eliminated. There is now abundant evidence, as reviewed by Kærn *et al.* [12], Raser and O'Shea [13], and Larson *et al.* [14], showing that even isogenic populations grown in homogeneous conditions are prone to significant phenotypic variability.

Previous theoretical studies of stochastic gene expression have focused on quantifying variation in mRNAs and proteins found in isogenic populations of cells, either through the calculation of moments of random copy number [15–18] or by finding expressions for the distributions themselves [19–22]. Recently, eigenfunction methods have been introduced as a theoretical and computational tool in modeling gene networks [23, 24]. The population density approach has previously been applied to the problem of stochastic gene expression without explicit consideration of the stochasticity of gene-state switching [19]. Pedraza and Paulsson studied the effect of non-exponential waiting times between transcriptional bursts on the stationary variance in protein abundance [25]. They found that protein noise was insensitive to the distribution of waiting time in the inactive gene state, in an instantaneous burst model. Other theoretical studies have focused on the control of noise by feedback [16, 18, 26], an aspect of stochastic gene expression that is omitted in this work.

In the case of constitutively active genes, mRNA and protein synthesis and decay are the key discrete, stochastic events leading to cell-to-cell variability in mRNA and protein levels [16, 17, 27], and models based on these events have indeed proven successful in prokaryotic organisms [7, 28]. However, a growing body of experimental evidence, both direct [21, 29, 30] and indirect [31–34], shows that many, if not most, higher eukaryotic genes themselves stochastically transition between transcriptionally active and transcriptionally inactive states. In this scenario, transcription occurs in *bursts*, and this introduces another significant (often dominant) source of variability. There are examples of such behavior in bacteria [28, 29], although other work indicates that this is not always the case [7]. Single-mRNA counting experiments in yeast [35] provide evidence for gene switching in the case of a transcriptionally regulated gene but not for constitutively active genes.

In many previous models of stochastic gene expression the processes of gene activation and inactivation have been modeled with first order kinetics [10, 11, 15, 22, 31, 36–40]. This corresponds to exponentially distributed gene-state dwell times. Because the initiation of transcription and gene inactivation by chromatin remodeling are known to involve several kinetic steps [37, 41–43], the assumption of exponential dwell times seems unlikely to be correct. The likelihood method of Suter *et al.* [44], applied to experimental data, allowed them to infer non-exponential dwell times in the inactive gene state in particular. Thus, by generalizing the gene-state switching model, a more realistic description of the physical system is possible. Our generalization allows us to determine if the first order switching model simply provides a crude but servicible approximation or if the steady-state mRNA distributions depend only weakly on the kinetic details of gene switching. Furthermore we can explore the extent to which temporal data in the form of the steady-state autocovariance function can be used to disambiguate candidate kinetic models.

II. MODEL AND RESULTS

We study a model for stochastic gene activation and inactivation, stochastic transcription, and molecular degradation of mRNA. This extends the first order gene-state kinetics of many previous models to more general gene-state switching models. The reaction scheme is summarized by the reaction diagram:

$$\underset{\text{dwell time } T_0}{\text{I}} \longleftrightarrow \underset{\text{dwell time } T_1}{\text{A}} \tag{1}$$

$$\text{A} \xrightarrow{\nu} \text{A} + \text{mRNA} \tag{2}$$

$$\text{mRNA} \xrightarrow{\delta} \emptyset \tag{3}$$

A gene of interest fluctuates randomly, Eq. (1), between an active state A, in which transcription occurs, and an inactive state I, in which transcription cannot occur. The phrase "transcriptional burst" refers to the synthesis of mRNA during active-gene

epochs separated by periods in which the gene is transcriptionally inactive.

In Eq. (1), T_0 and T_1 are random variables with probability densities f_{T_0} and f_{T_1} respectively. When the gene is in the active state, Eq. (2), the synthesis of mRNA molecules is governed by a Poisson process, with a probability per unit time of a synthesis event equal to ν. Each mRNA molecule is stochastically degraded, Eq. (3), with probabilities per unit time, per molecule, of δ.

There are two discrete state variables in the stochastic system, Eqs. (1)–(3): $M(t)$, the random number of mRNA molecules at time t; $Z(t)$, a dichotomous random variable, where $Z(t) = 0$ if the gene is in the inactive (off) state I, and $Z(t) = 1$ if the DNA is in the active (on) state A at time t. In addition there are two continuous random variables: S_0, the time since entering the inactive state, and S_1, the time since entering the active state. In the corresponding, analogous, continuous random mRNA model, motivated below, the discrete gene-state variable, Z, is the same, and we refer to the continuous mRNA random variable as $Y(t)$. Initial conditions for this stochastic process are irrelevant since we will only consider the steady state of this system.

A. Link between the discrete and continuous mRNA models

As pointed out by Iyer-Biswas *et al.* [38], a probability mass function for the discrete random mRNA copy number, M, can be derived from the probability density function for the continuous random mRNA copy number, Y, in an analogous continuous mRNA problem. A more general use of this connection will be a theme of this paper. We emphasize that, in the present paper, the continuous mRNA problem does not serve as an approximate model in any sense. Rather, it serves a key mathematical role in the derivation of exact mRNA distributions in the fully discrete problem.

In both problems (continuous and discrete mRNA), the gene-state variable switches stochastically between the active and inactive states. The continuous mRNA model has previously been studied in the case of exponentially distributed gene-state dwell times [36]. In this model, transcription and degradation of mRNA are deterministic processes for any given history of the dichotomous random gene-state function, $Z(t)$; the continuous mRNA copy number, $Y(t)$, is random only because the gene-state function is random.

The random differential equation for the continuous mRNA copy number is

$$\frac{dY}{dt} = \nu\, Z(t) - \delta\, Y, \tag{4}$$

where $Z(t)$ is the dichotomous random gene-state function. For any given gene-state history function $Z(t)$, Eq. (4) can be solved analytically. Given the specific initial condition that $Y(t = 0) = 0$,

$$Y(t) = \int_0^t \nu\, Z(t') \exp\Big(- (t - t')\, \delta \Big) dt'. \tag{5}$$

We define $f_Y(y, t)$ to be the probability density function for $Y(t)$, i.e. $f_Y(y, t)\, dy = \Pr\{Y(t) \in (y, y + dy)\}$. It is worth noting that the continuous mRNA copy number cannot exceed ν/δ, provided that $Y(t = 0) \in [0, \nu/\delta]$, because ν/δ is the steady-state value that Y approaches when the gene is fixed in the active state. Consequently, $Y(t) \in [0, \nu/\delta]$ for all t.

In the discrete mRNA version of the reaction scheme depicted in Eqs. (1)–(3) the synthesis and degradation events occur at points in time. Synthesis occurs with probability per unit time, $\nu\, Z(t)$, and each synthesis event increases the mRNA copy number by 1. Degradation occurs with probability per unit time, per molecule, δ, and each degradation event decreases the mRNA copy number by 1. The solution for the probability mass function for the random mRNA copy number $M(t)$ in this classical birth and death process [45], with any given $Z(t)$, is a Poisson probability mass function at each point in time with a mean given by $Y(t)$, as defined in Eq. (5), i.e., the solution to the corresponding continuous mRNA problem above. This can be derived by first writing down the master equation for $p_{M|Z(\cdot)}(m, t; Z(\cdot))$, which can be solved by use of the probability generating function to give $p_{M|Z(\cdot)}(m, t; Z(\cdot)) = e^{-Y(t)} \frac{Y(t)^m}{m!}$ with $Y(t)$ given by Eq. (5). Thus, the unconditioned probability mass function for $M(t)$ is given by

$$\begin{aligned} p_M(m, t) &= \int_0^{\nu/\delta} p_{M|Y}(m; y)\, f_Y(y, t)\, dy \\ &= \int_0^{\nu/\delta} \frac{y^m}{m!} e^{-y} f_Y(y, t)\, dy, \quad \text{for } m \in \mathbb{Z} \geq 0, \end{aligned} \tag{6}$$

in which $f_Y(y, t)$ depends on the initial distribution of S_0, S_1, and Z. Knowledge of $f_Y(y, t)$ for the continuous mRNA problem fully specifies the probability mass function for mRNA copy number in the discrete problem. Details of this representation of $p_M(m, t)$ were given by Iyer-Biswas [46]. Note that Eq. (6) reflects the fact that, in the fully discrete problem, the random mRNA copy number M is not bounded, even though its continuous analog Y is bounded by ν/δ.

Before studying the arbitrary dwell times model in detail, we review results for the simpler case of exponential dwell times. Karmakar and Bose [36] found the steady-state solution for $f_Y^\infty(y)$, in the case of exponential dwell times $f_{T_0}(t_0) = \lambda e^{-\lambda t_0}$ and

$f_{T_1}(t_1) = \gamma e^{-\gamma t_1}$, to be a beta density function with shape parameters λ/δ and γ/δ, and scale parameter ν/δ:

$$f_Y^{\infty}(y) = \frac{1}{\nu/\delta}\,\frac{\Gamma(\lambda/\delta\,+\,\gamma/\delta)}{\Gamma(\lambda/\delta)\,\Gamma(\gamma/\delta)}\left(\frac{y}{\nu/\delta}\right)^{\lambda/\delta-1}\left(1-\frac{y}{\nu/\delta}\right)^{\gamma/\delta-1}, \quad \text{for } y \in [0, \nu/\delta]. \tag{7}$$

With this solution for $f_Y^{\infty}(y)$ in hand, one can find the corresponding steady-state probability mass function for the discrete mRNA problem, $f_M^{\infty}(m)$, by substituting the explicit expression for $f_Y^{\infty}(y)$, Eq. (7), into Eq. (6) to obtain

$$p_M^{\infty}(m) = \int_0^{\nu/\delta} \frac{y^m}{m!}\,e^{-y}\,\frac{1}{\nu/\delta}\,\frac{\Gamma(\lambda/\delta\,+\,\gamma/\delta)}{\Gamma(\lambda/\delta)\,\Gamma(\gamma/\delta)}\left(\frac{y}{\nu/\delta}\right)^{\lambda/\delta-1}\left(1-\frac{y}{\nu/\delta}\right)^{\gamma/\delta-1} dy. \tag{8}$$

Eq. (8) is a Poisson-beta distribution [47], which can also be written in terms of the confluent hypergeometric function ${}_1F_1$, as originally stated by Raj *et al.* [11] and also later verified independently by Shahrezaei and Swain [22]. In a *tour de force*, Iyer-Biswas *et al.* [38] derived an analytical expression for the corresponding time-dependent probability generating function. The steady-state probability generating function was obtained previously by Peccoud and Ycart [15].

The remainder of this paper explores many powerful consequences of Eq. (6).

B. Formulation of the problem with arbitrary gene-state dwell times

Following the work of Iyer-Biswas *et al.* [38], we exploit the Poisson connection stated in Eq. (6) to solve for the distribution of discrete mRNA copy number M by first solving for the density of the continuous random variable Y. In the gene switching problem with non-exponential dwell times, the probability per unit time of switching out of a state depends on the time since entering that state. Consider the off state for example. If the dwell times T_0 in the off state are independent random variables with density function $f_{T_0}(t_0)$ and $S_0(t)$ is the random elapsed time since entering the off state, then $\Pr\{\text{ switch on in } (t, t+dt) \,|\, S_0(t) = s_0\} = h_{T_0}(s_0)\,dt$, where h_{T_0} is the hazard function for the random variable T_0; $h_{T_0}(s_0) = f_{T_0}(s_0)/\tilde{F}_{T_0}(s_0)$ and $\tilde{F}_{T_0}(s_0)$ is the complementary cumulative distribution function. Note that when a gene is in the off state $\frac{dS_0}{dt} = 1$ and similarly for a gene in the on state $\frac{dS_1}{dt} = 1$.

The evolution of population density in the continuous mRNA problem is governed by a conservation of probability. The partial differential equations for the continuous random mRNA copy number in the inactive and active states are

$$\begin{aligned}
\frac{\partial}{\partial t} f_{Y,Z,S_0}(y,0,s_0,t) &= -\frac{\partial}{\partial y}\left[-\delta\, y\, f_{Y,Z,S_0}(y,0,s_0,t)\right] \\
&\quad -\frac{\partial}{\partial s_0}\left[1\cdot f_{Y,Z,S_0}(y,0,s_0,t)\right] \\
&\quad -h_{T_0}(s_0)\, f_{Y,Z,S_0}(y,0,s_0,t), \qquad (9)\\
\frac{\partial}{\partial t} f_{Y,Z,S_1}(y,1,s_1,t) &= -\frac{\partial}{\partial y}\left[(\nu-\delta\, y)\, f_{Y,Z,S_1}(y,1,s_1,t)\right] \\
&\quad -\frac{\partial}{\partial s_1}\left[1\cdot f_{Y,Z,S_1}(y,1,s_1,t)\right] \\
&\quad -h_{T_1}(s_1)\, f_{Y,Z,S_1}(y,1,s_1,t). \qquad (10)
\end{aligned}$$

The s variables record the elapsed time since the gene switched states. The first two terms on the right-hand side of each equation correspond to the divergence of the probability flux. The synthesis and degradation dynamics give rise to an advective probability flux in the first term in each equation. The evolution of elapsed time since entering the inactive and active states, with unit velocity, gives rise to the advective probability flux in the second term in each equation. The third term in each equation represents a sink due to the gene switching between its inactive and active states. The dynamics of the two probability densities are coupled only through the boundary conditions at $s = 0$ which follow.

At the instant a gene in the off state switches on, the elapsed time in the on state, S_1, is equal to zero, and similarly for a switch in the opposite direction. A consequence is a probability flux in the s_0 and s_1 directions at $s_0 = 0$ and $s_1 = 0$, respectively. This flux has contribution from members of the population with elapsed times in the state being exited:

$$f_{Y,Z,S_0}(y,0,s_0=0,t) = \int_0^{\infty} h_{T_1}(s_1)\, f_{Y,Z,S_1}(y,1,s_1,t)\, ds_1, \tag{11}$$

$$f_{Y,Z,S_1}(y,1,s_1=0,t) = \int_0^{\infty} h_{T_0}(s_0)\, f_{Y,Z,S_0}(y,0,s_0,t)\, ds_0. \tag{12}$$

We solve for the steady state of this system, Eqs. (9)–(12), numerically. To this end, we write the equations in dimensionless form. Time and elapsed time in a particular gene state are scaled by the rate at which mRNA is degraded: $\tau = \delta t$ and $\sigma = \delta s_i$; and the continuous mRNA variable is scaled by its maximum value: $x = \frac{\delta}{\nu} y$. The subscript on σ is dropped and σ is understood to be the elapsed time in whichever gene state is relevant. After introducing two scaled densities, $v(x,\sigma) = \frac{\nu}{\delta} f_{Y,Z,S_0}(\frac{\nu}{\delta}x, 0, \frac{\sigma}{\delta}, t \to \infty)$ and $u(x,\sigma) = \frac{\nu}{\delta} f_{Y,Z,S_1}(\frac{\nu}{\delta}x, 1, \frac{\sigma}{\delta}, t \to \infty)$, as well as scaled hazard functions, $h_{\mathcal{T}_0}(\sigma) = h_{T_0}(\frac{\sigma}{\delta})/\delta$ and $h_{\mathcal{T}_1}(\sigma) = h_{T_1}(\frac{\sigma}{\delta})/\delta$, the equations become

$$\frac{\partial v}{\partial \sigma} - x\frac{\partial v}{\partial x} = (1 - h_{\mathcal{T}_0}(\sigma))\, v, \tag{13}$$

$$\frac{\partial u}{\partial \sigma} + (1-x)\frac{\partial u}{\partial x} = (1 - h_{\mathcal{T}_1}(\sigma))\, u, \tag{14}$$

$$v(x,0) = \int_0^\infty h_{\mathcal{T}_1}(\sigma)\, u(x,\sigma)\, d\sigma, \tag{15}$$

$$u(x,0) = \int_0^\infty h_{\mathcal{T}_0}(\sigma)\, v(x,\sigma)\, d\sigma. \tag{16}$$

Each of Eqs. (13) and (14) are linear, first order partial differential equations which can be solved analytically within $\sigma > 0$ by the method of characteristics in terms of the function values at $\sigma = 0$ with σ parameterizing the characteristics. This problem can be reduced to a pair of integral equations for the unknowns immediately after a gene switch when $\sigma = 0$, $v_0(x) \equiv v(x,0)$ and $u_0(x) \equiv u(x,0)$. The details appear in Appendix A. The result is the eigenfunction problem

$$v_0(x) = \frac{1}{1-x}\int_0^x f_{\mathcal{T}_1}\left(\log\left(\frac{1-x_0}{1-x}\right)\right) u_0(x_0)\, dx_0, \tag{17}$$

$$u_0(x) = \frac{1}{x}\int_x^1 f_{\mathcal{T}_0}\left(\log\left(\frac{x_0}{x}\right)\right) v_0(x_0)\, dx_0. \tag{18}$$

Solving for $u_0(x)$ and $v_0(x)$ allows one to compute the solution, $u(x,\sigma)$ and $v(x,\sigma)$. Also, the steady-state marginal mRNA densities, $v_{\mathrm{ss}}(x) = \int_0^\infty v(x,\sigma)\, d\sigma$ and $u_{\mathrm{ss}}(x) = \int_0^\infty u(x,\sigma)\, d\sigma$, can be extracted from $v_0(x)$ and $u_0(x)$ by computing the integrals (see Appendix A)

$$v_{\mathrm{ss}}(x) = \frac{1}{x}\int_x^1 \tilde{F}_{\mathcal{T}_0}\left(\log\left(\frac{x_0}{x}\right)\right) v_0(x_0)\, dx_0, \tag{19}$$

$$u_{\mathrm{ss}}(x) = \frac{1}{1-x}\int_0^x \tilde{F}_{\mathcal{T}_1}\left(\log\left(\frac{1-x_0}{1-x}\right)\right) u_0(x_0)\, dx_0. \tag{20}$$

Note that Eqs. (17) and (18) are homogeneous and to obtain a unique solution, the total probability condition $\int_0^1 (u_{\mathrm{ss}}(x) + v_{\mathrm{ss}}(x))\, dx = 1$ must be used.

For all but the most trivial dwell time densities, the integral equations, Eqs. (17) and (18), must be solved numerically. We discretized the integrals with the trapezoidal rule to form a finite linear system. The discretized equations, with discrete linear operator acting on the discretized samples of $v_0(x)$ and $u_0(x)$, can be viewed as an eigenvector problem with eigenvalue that should be equal to one. However, the relevant matrix does not have an eigenvalue of exactly one. We observe empirically that the largest, in modulus, of its eigenvalues approaches one as the number of grid points increases. We take the eigenvector corresponding to the eigenvalue nearest to one as the solution to the discretized problem. This eigenvector is computed using the inverse power method. Alternatively, a least squares minimizer to the eigenvector problem can be used. Both methods result in numerical schemes that approach the solution of the continuous problem as the grid is refined. This claim can be explicitly verified by comparison to the known steady-state solution, Eq. (7), in the case of exponential dwell times.

The steady-state solution of the continuous mRNA problem for arbitrary dwell time densities allows one to study the steady-state solution of the discrete mRNA problem with arbitrary dwell time densities. The same distributions could be estimated with Monte Carlo simulations, but the above procedure is considerably faster for a given level of accuracy.

C. Model solutions

We use two different gene-state dwell time models to illustrate solutions of the general model described above. A common gene-state switching model is one in which the gene-state dwell times are exponentially distributed, $f_{T_0}(t_0) = \lambda e^{-\lambda t_0}$ and $f_{T_1}(t_1) = \gamma e^{-\gamma t_1}$. This corresponds to simple, first order kinetics of the gene-state switching. A second gene-state switching model will be considered in which the dwell times in the active and inactive states have a gamma distribution, $f_{T_i}(t_i) = t_i^{k_i - 1} e^{-t_i/\theta_i}/(\Gamma(k_i)\theta_i^{k_i})$.

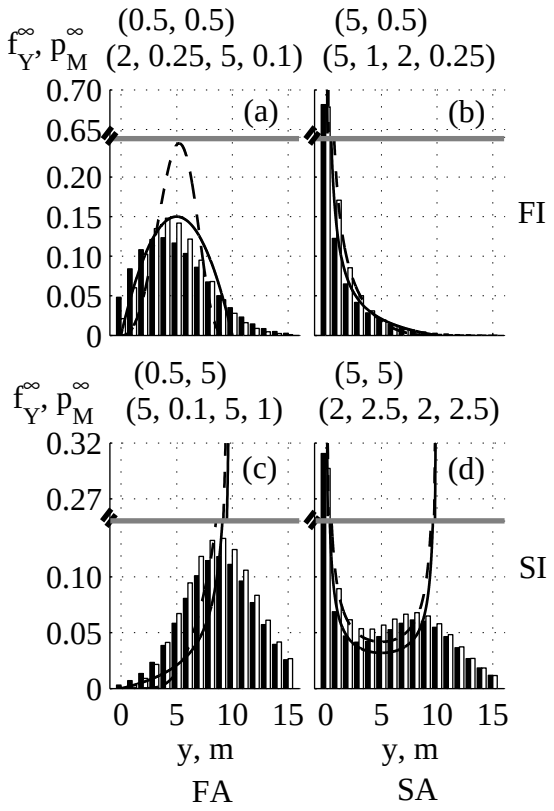

FIG. 1. Steady-state mRNA distributions for four sets of parameters of the exponential dwell times model and the gamma dwell times model: $(\delta/\lambda, \delta/\gamma)$ and $(k_0, \theta_0\delta, k_1, \theta_1\delta)$ with $\nu/\delta = 10$ for both models. The probability mass function for the random mRNA copy number in the fully discrete problem for the exponential dwell times model (black bars) and the gamma dwell times model (white bars) are shown along with the *parent* probability density function from the continuous problem for the exponential dwell times model (solid) and the gamma dwell time model (dashed). The left column corresponds to fast activation (FA) and the right to slow activation (SA), while the first row corresponds to fast inactivation (FI) and the second to slow inactivation (SI).

The shape parameter, k, and the scale parameter, θ, take on positive values. The exponential dwell times model is recovered when $k_0 = 1 = k_1$. For integer k, such a distribution can be thought of as a sum of k independent exponentially distributed random variables each with a mean of θ. If gene activation and inactivation were governed by numerous small steps, with similar rates, that are nearly irreversible, the gamma distribution for the gene-state dwell times would be a good approximation. Additionally, the motivation for such a dwell time distribution is to account for a refractory period in each state. Experimental evidence [44] suggests that there exists such a refractory period, at least in the inactive gene state.

Fig. 1 displays the steady-state probability mass function, $p_M^\infty(m)$, for random mRNA copy number for four sets of parameter values for each of the two models along with the *parent* probability density function for the corresponding continuous mRNA problem, $f_Y^\infty(y)$. We refer to $f_Y^\infty(y)$ as the parent probability density function because $f_Y^\infty(y)$ can be thought of as giving rise to the discrete mRNA distribution though Eq. (6). All parameters are reported in ratio to the mRNA degradation rate, δ, and are displayed above each panel. In all cases, $\nu/\delta = 10$, which represents a moderately fast transcription rate. The upper row has parameters that correspond to fast inactivation of the gene, while the lower row has parameters for slow gene inactivation. The left column corresponds to fast activation and the right to slow activation. As a point of reference, the parameter regime of slow gene activation (small λ/δ), fast gene inactivation (large γ/δ), and fast transcription (large ν/δ) corresponds to measurements of eukaryotic cells by Raj *et al.* [21]; slow gene activation, fast gene inactivation, and slow transcription corresponds to measurements in bacteria by Golding *et al.* [29].

Note that for slow gene inactivation rate (lower row), the parent beta density function shows a sharp peak at ν/δ. In the discrete model, by contrast, there cannot be a sharp peak at ν/δ. The sharp peak in the parent beta density function is smoothed because the probability mass function is a weighed average of Poisson distributions. The physical basis for this smoothing can be thought of as Poisson variation about a mean of ν/δ during prolonged active-gene epochs.

Within each panel of Fig. 1, the exponential dwell times model is compared with the gamma dwell times model to observe the effect of changes to the dwell times model. The mean dwell time in the active and inactive gene states are selected to be same in each case. The probability mass functions nonetheless appear qualitatively similar. Later in this paper, we address the question as to what extent can experimental data be used to reject the model of exponential dwell times in favor of the gamma distributed dwell times.

As another demonstration of the generality of our approach, we studied the model of Tang [48] in which the inactive gene state is considered to have two sequential steps with exponentially distributed durations and means $1/\kappa$ and $1/\lambda$. The inactive dwell time density is given by $f_{T_0}(t_0) = \frac{\kappa\lambda}{\lambda-\kappa}\left(e^{-\kappa t_0} - e^{-\lambda t_0}\right)$. The active dwell time density is exponential, $f_{T_1}(t_1) = \gamma e^{-\gamma t_1}$. Suter *et al.* [44] showed that maximum likelihood estimates of gene-state paths are consistent with this model and these authors inferred parameter values for various genes. Typical values from the Suter *et al.* [44] study were $\kappa = 0.04$/min, $\lambda = 0.04$/min, $\gamma = 0.1$/min, $\nu = 0.25$/min, and $\delta = 0.01$/min. With these values, the steady-state copy number probabilities are $p_M(m) = 0.07, 0.13, 0.15, 0.14, 0.13, 0.10, 0.08, 0.06, 0.04, 0.03, 0.02$ for $m = 0, \ldots, 10$. The mean and standard derivation are 4.2 and 3.1 respectively. The calculations of Tang give time-dependent moments, while our method gives the steady-state copy-number distribution. Our computed steady-state moments agree with those of Tang.

D. Ambiguity in the steady-state distributions

Although exponentially distributed gene-state dwell times are commonly used (cf. references above), some experimental evidence (see [44]), suggests that the switching dynamics are more complicated. The numerous steps involved in transcriptional regulation are discussed by Pedraza and Paulsson [25]. Given that most experimental techniques available today observe the number of mRNA molecules and not gene state, the natural question arises as to how data involving mRNA copy number can reveal details of the gene-state switching. It is worth noting that when activation and inactivation are very fast relative to the decay rate of mRNA molecules, a Poisson probability mass function will be a good approximation to the distribution of mRNA copy numbers regardless of the dwell time distributions.

How well do the data constrain the model? To investigate this question, we generated data from a gene-state dwell times model consisting of gamma densities with parameters $k_0 = 2, \theta_0 = 2.5, k_1 = 2, \theta_0 = 2.5$. mRNA copy number data was generated from this switching model by sampling the computed probability mass function from the numerical method described above for 1,000 simulated cells. The data are displayed in Fig. 2A along with the underlying distribution from which they are sampled. Maximum likelihood parameters and likelihood of the data for the model with exponential gene-state dwell times were computed. The probability mass function with maximum likelihood parameters for the exponential dwell times model is also shown in Fig. 2A. We found that the likelihood of the data generated from the gamma dwell times model fell in the middle, at the 0.77 quantile, of the distribution of likelihoods, with 1,000 simulated data sets, for data generated by the exponential dwell times model with maximum likelihood parameters. This is shown in Fig. 2B. Thus, the exponential dwell times model cannot be rejected even though the true dwell time densities are appreciably different than the implied densities. Fig. 2C displays the contours of the likelihood function near the maximum likelihood parameters.

The observation that the exponential dwell times model cannot be rejected from the steady-state mRNA distribution is not specific to the parameters selected for the gamma densities. We repeated the procedure above for a number of different combinations of parameters: active dwell time gamma density parameters $k_1 = 2$ and $\theta_1\delta = 0.25, 2.5$ along with inactive dwell time gamma density parameters $k_0 = 5$ and $\theta_0\delta = 0.1, 1, 0.02, 0.2$ or $k_0 = 10$ and $\theta_0\delta = 0.1, 1$. Two transcription rates were used: $\nu/\delta = 10, 100$. For each of the 24 cases, 1,000 cells were simulated with a total of 25 datasets. In all cases, the exponential dwell times model could not be rejected. The -log-likelihood of the data generated from the gamma dwell times model always fell in the distribution of -log-likelihoods with data generated from exponential dwell times model (with the maximum likelihood parameters) below the 0.80 quantile. Additionally, in every case the mean dwell times in the active and inactive states were underestimated by the maximum likelihood exponential dwell times model, sometimes a factor of 5 lower, despite a close approximation of the steady-state mRNA distribution. The same procedure was applied to data generated from the model of Tang [48] using parameters reported in Suter *et al.* [44]. We again found that the exponential dwell times model could not be rejected.

We have shown that an experimental estimate of the steady-state mRNA distribution is insufficient to reveal non-exponential dwell times of the gene switching based on a realistic number of 1,000 cells. We obtained similar results in Monte Carlo simulations with 10,000 cells. In selecting k_0 equal to 5 or 10, we have shown that even extreme departures from exponential dwell times are not revealed in the steady-state mRNA copy number data. Consequently, we asked if measuring and fitting the autocovariance function could be used to discriminate among models.

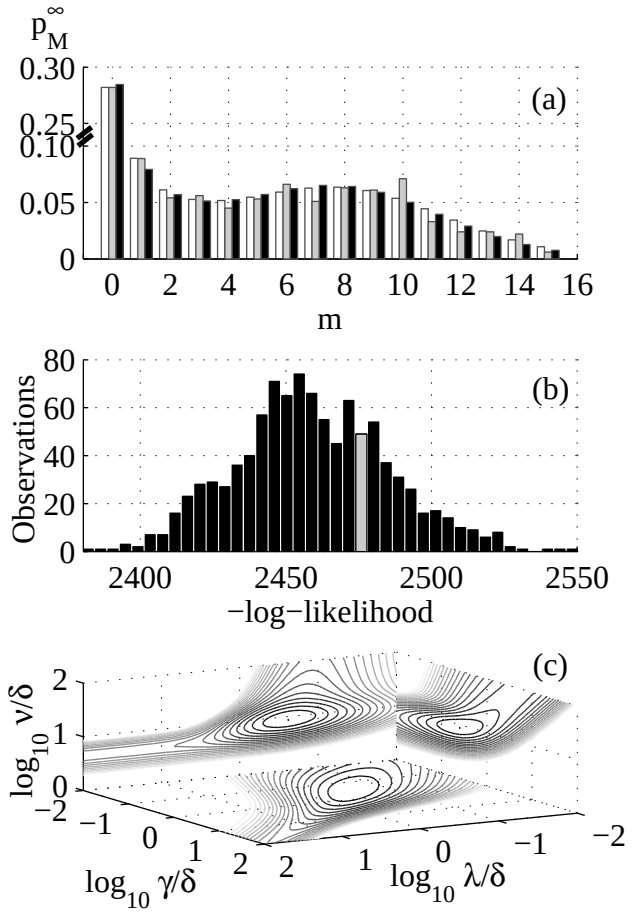

FIG. 2. Details of the parameter estimation process for the parameters: $(k_0, \theta_0\delta, k_1, \theta_1\delta, \delta/\nu) = (2.0, 2.5, 2.0, 2.5, 0.10)$, $\left((\delta/\lambda)^{ML}, (\delta/\gamma)^{ML}, (\delta/\nu)^{ML}\right) = (4.0, 4.0, 0.11)$. (a) The white bars are the underlying distribution, the gray bars are a histogram of the simulated data from 1,000 cells, and the black bars are the probability mass function for the maximum likelihood parameters of the exponential dwell times model. (b) Histogram of the -log-likelihood values of 1,000 data sets, each consisting of simulated measurements of mRNA copy numbers in 1,000 cells. The bin containing the likelihood of the data above (2475) is identified. (c) Contours of log-likelihood as a function of model parameters. Slices of the 3D function occur at $\lambda/\delta = (\lambda/\delta)^{ML}$, $\gamma/\delta = (\gamma/\delta)^{ML}$, and $\nu/\delta = (\nu/\delta)^{ML}$. The largest contour value is -2550 and the contour values have a uniform spacing of 130.

E. Autocovariance function for mRNA copy number

In Appendix B we present a derivation of a formula for the steady-state autocovariance function for mRNA copy number, $\Phi_{MM}(\tau) \equiv \mathbb{E}\left[M(t)\, M(t+\tau)\right] - \mu_M^2$. The autocovariance is a function of only the delay, τ, since the process is assumed to be stationary. The autocovariance function is the sum of two terms. One is a simple exponential with a decay rate equal to the rate of decay of mRNA molecules, and the other is a convolution of the same exponential function with the autocovariance function, $\Phi_{ZZ}(\tau)$, of the random gene-state function $Z(t)$:

$$\Phi_{MM}(\tau) = \frac{\nu}{\delta}\,\mu_Z\, e^{-\delta|\tau|} + \frac{\nu^2}{2\delta}\int_{-\infty}^{\infty} e^{-\delta|s|}\,\Phi_{ZZ}(\tau - s)\,ds, \tag{21}$$

where $\mu_Z = \frac{\mu_{T_1}}{\mu_{T_0}+\mu_{T_1}}$ is the steady-state mean of $Z(t)$, i.e., the fraction of time that the gene is in the active state. The power spectrum of the mRNA autocovariance function is related to the power spectrum of gene-state autocovariance function by

$$\hat{\Phi}_{MM}(\omega) = \frac{2\nu}{\delta^2+\omega^2}\left(\mu_Z + \frac{\nu}{2}\hat{\Phi}_{ZZ}(\omega)\right). \tag{22}$$

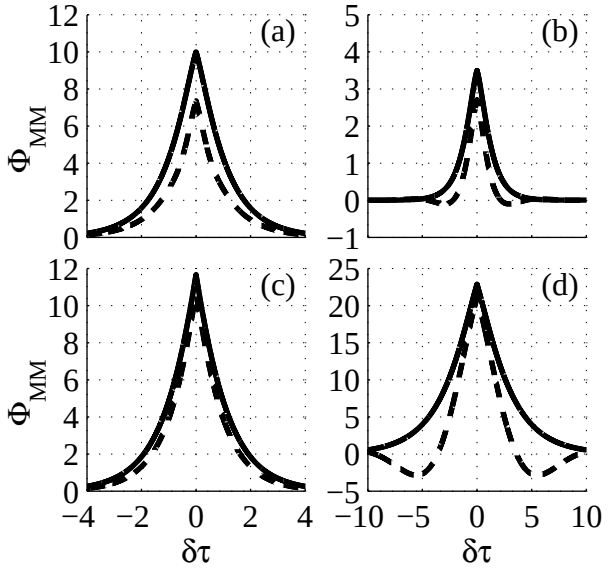

FIG. 3. mRNA autocovariance functions for the exponential dwell times model (solid) and the gamma dwell times model (dashed) for the four parameter sets in Fig. 1.

$\Phi_{ZZ}(\tau)$ can be computed from the dwell time densities by

$$\Phi_{ZZ}(\tau) = \mu_Z(1-\mu_Z) - \frac{1}{\mu_{T_0}+\mu_{T_1}} \int_0^{|\tau|} (|\tau| - t) \int_{-\infty}^{\infty} \frac{e^{i\omega t}}{2\pi} \frac{(1-\hat{f}_{T_0}(\omega))(1-\hat{f}_{T_1}(\omega))}{1-\hat{f}_{T_0}(\omega)\hat{f}_{T_1}(\omega)}\, d\omega\, dt. \tag{23}$$

An equivalent formula was presented previously in the context of modeling air conditioning loads on electrical power systems [49]. Its derivation appears in Appendix B. Note that $\Phi_{ZZ}(0) = \mu_Z(1-\mu_Z)$, which is the variance in Z. The beauty of Eqs. (21) and (23) is their generality. For any desired dwell time densities of practical interest, one can compute the autocovariance of the gene state and the mRNA copy numbers.

For the particular gene switching scheme of exponential dwell times, the autocovariance function of $M(t)$ evaluates to a difference of two exponentials. One rate constant is the rate at which the random gene-state function, $Z(t)$, relaxes to its steady-state distribution, $(\lambda+\gamma)$, and the other is the rate of decay of mRNA molecules, δ. The formula is

$$\begin{aligned} \Phi_{MM}(\tau) = \quad & \mu_M \left(1 + \frac{\nu\gamma}{(\lambda+\gamma)^2 - \delta^2}\right) e^{-\delta|\tau|} \\ & - \mu_M^2 \frac{\gamma}{\lambda} \frac{\delta^2}{(\lambda+\gamma)^2 - \delta^2}\, e^{-(\lambda+\gamma)|\tau|}. \end{aligned} \tag{24}$$

Here $\mu_M = \frac{\nu}{\delta}\mu_Z = \frac{\nu}{\delta}\frac{\lambda}{\lambda+\gamma}$.

Fig. 3 shows plots of $\Phi_{MM}(\tau)$ on a dimensionless time axis in which the abscissa is $\delta\tau$. Each set of parameters for the exponentially and gamma distributed dwell times from Fig. 1 is shown. The monotonic nature of the decay of $\Phi_{MM}(\tau)$ is generic for the exponential dwell times model. It can be shown that in the exponential dwell times model, $\Phi_{MM}(\tau)$ cannot have an undershoot (cannot take on negative values). Negative autocovariance in the gene-state variable is possible with gamma distributed dwell times, as seen in panels B and D.

The qualitative differences between the autocovariance functions from the exponential and gamma distributed dwell time models suggest that the autocovariance function may be useful for discriminating among gene-state dwell time models.

F. Model discrimination using the autocovariance function

Recent experimental advances [29, 44, 50–52], have enabled counting of individual molecules over time. This permits the use of mRNA measurements over time to identify gene-state switching models. Suter *et al.* [44] demonstrated that temporal measurements and approximate likelihood inference methods can be used to identify non-exponential dwell time distributions.

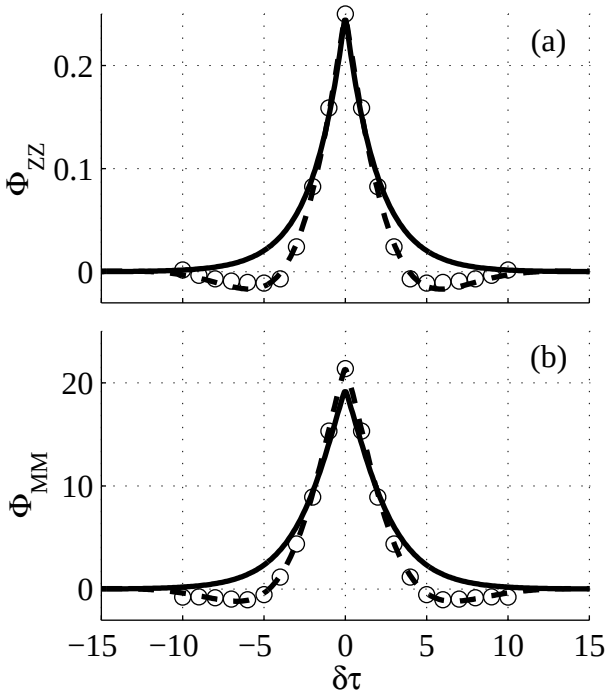

FIG. 4. Steady-state autocovariance functions. (a) $\Phi_{ZZ}(\tau)$ for maximum likelihood parameters of the exponential dwell times model (solid) with $\Phi_{ZZ}(\tau)$ of the model underlying the simulated data (dashed). The open circles are an estimate of the autocovariance function from a Monte Carlo simulation. (b) Same for $\Phi_{MM}(\tau)$.

Attempts to use transient temporal data following the induction of transcription would likely be unsuccessful due to the difficulty in obtaining the sufficient data during a single transient window per cell. A viable alternative appears to be the use of correlations in mRNA copy number, which includes temporal measurements, but not transient effects. Fig. 4 shows the gene-state autocovariance function, $\Phi_{ZZ}(\tau)$, and the mRNA autocovariance function, $\Phi_{MM}(\tau)$, for the same gamma distributed dwell times model and parameters as in Fig. 2 (dashed lines). The solid lines are for the exponential dwell times model with maximum likelihood parameters. The open circles are an estimate of the autocovariance functions from a Monte Carlo simulation, in which temporal data is sampled once every $1/\delta$ units of time for a total of 1,000 samples. The standard *unbiased* estimator of autocovariance with estimated mean was used. The early portion of the Monte Carlo simulation was discarded so that the autocovariance is estimated once the stochastic process was stationary. Autocovariance estimates are shown only when at least 100 samples, for the particular τ, are available.

Near $\tau = 0$, the autocovariance functions are in close agreement, which reflects the fact that the steady-state distributions have a similar variance. However, the autocovariance functions disagree for larger τ and the autocovariance for the true model displays the distinctive feature of negative values, which the exponential model is not capable of producing. The estimates of the autocovariance function from the simulated temporal data agree more closely with the true model, suggesting that these data could reveal the underlying gene switching dynamics.

For a broad range of parameters in the gamma dwell times model, the autocovariance function was sampled at a frequency 10 times its fastest time scale for a duration of 10 times its slowest time scale of the generative model. A least squares fitting procedure was used to identify parameters in the exponential dwell time model. The total size of the square errors relative to the magnitude of the autocovariance samples was used as a measure of the closeness of the fit. We observed that a poor fit was obtained only when the autocovariance of the underlying model had negative values. Consequently, anticorrelation in mRNA copy number is necessary to reject the exponential dwell times model.

It is difficult to fully characterize the model parameters for which negative autocovariance is predominant. For concreteness, we take $k_0 = 5$ and $k_1 = 2$ and measure the prevalence of negative autocovariance as the absolute value of the most negative value of Φ_{MM} divided by $\Phi_{MM}(0)$. The contours of this index are plotted in Fig. 5 against the mean dwell times in each of the gene states. When either dwell time is short or long, the autocovariance function does not go negative. Negative autocovariance

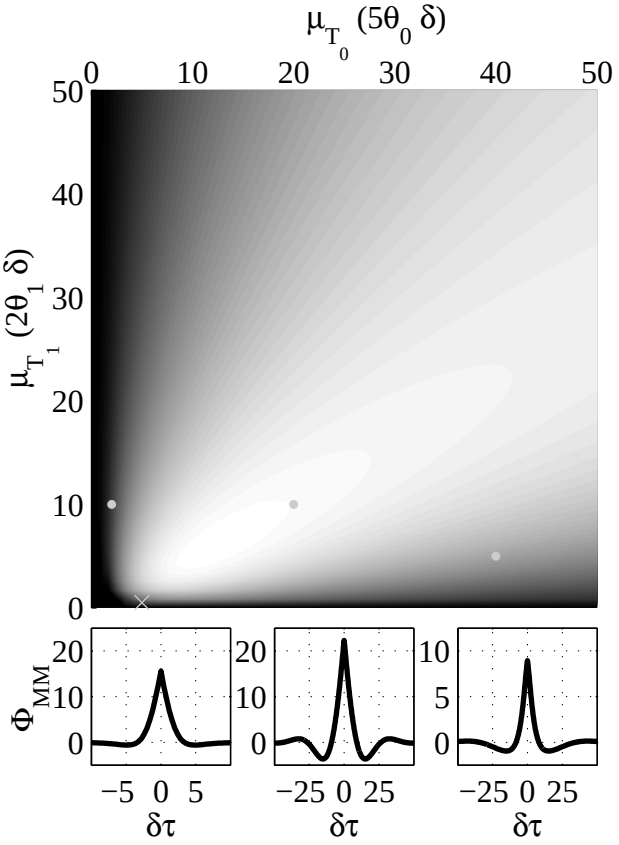

FIG. 5. Contour plot of the extreme negative value of the autocovariance function, Φ_{MM}, as a fraction of its value at zero in the gamma dwell times model with $k_0 = 5$ and $k_1 = 2$ over a range of mean dwell times in active and inactive gene states. The contours range from 0% (black) to 16% (white) in increments of 0.3%. Shown below are the autocovariance functions with parameters given by the corresponding dot above. The cross corresponds to the autocovariance function from Fig. 3b.

results from a balance between the duration in the inactive and active states and a low frequency switching between the two. The prevalence of negative autocovariance increases with the transcription rate, ν, and larger delays in either gene state (larger k_0 or k_1)

Negative mRNA autocovariance is possible in the model of Tang [48]. It is most prevalent when the mean durations of the two steps in the inactive states are equal ($1/\kappa = 1/\lambda$), which is equivalent to the gamma distributed dwell times model with $k_0 = 2$ and $k_1 = 1$. For the parameters inferred by Suter *et al.* [44], significant negative mRNA autocovariance does not occur in the model. In that study, transcriptional bursts were observed and significant negative mRNA autocovariance does not occur with bursting switching dynamics.

III. DISCUSSION

Raj *et al.* [11] measured the distribution of random mRNA copy number, for several different transcripts, over a population of homogeneous isogenic eukaryotic cells in culture. We have analyzed in this paper an extension of the stochastic transcription model that they used to explain and quantify their experimental data. In the model, the gene of interest fluctuates stochastically between an inactive state, in which transcription cannot occur, and an active state in which discrete transcription events occur; the individual mRNA molecules are degraded stochastically in an independent manner. The generalization in our model is that the dwell times in each state can be drawn from any given distributions. Our method provides a derivation of the distribution of random mRNA copy number that is based on the connection between the discrete mRNA random variable and the analogous, continuous mRNA random variable.

We exploited the Poisson connection between the continuous and discrete models to develop fast numerical methods for computing steady-state distributions for the discrete random mRNA copy number. In addition, we showed how this connection between continuous and discrete models, implicit in our calculation of $p_M^\infty(m)$, can be used to facilitate parameter inference by

maximum likelihood methods and model testing by likelihood methods.

Our results show that the steady-state mRNA distribution depends only weakly on the distributions of gene-state dwell times. This may explain the utility and robustness of the exponential dwell times model.

We presented analytic expressions for the gene-state and mRNA autocovariance functions. We showed, by example, that measurement of the autocovariance function for mRNA copy number in some parameter regimes could be used to discriminate among gene-state switching models that give indistinguishable steady-state mRNA distributions.

For the inferred parameters measured by Suter *et al.* [44] for various genes, the non-exponential dwell times model unfortunately does not exhibit negative autocovariance. This motivates future study of additional methods to distinguish gene-state switching models. One alternative is to fit the probability mass function $p^{\infty}_{M,Z=0}(m)$ and $p^{\infty}_{M,Z=1}(m)$ separately. This would require an experimental measurement of the gene state, which is currently available [51, 53]. It is likely that the most promising approach to model identification will involve computing the maximum likelihood of temporal sequences of experimentally measured mRNA copy numbers in the context of various different assumed densities for the dwell times. For each candidate model the likelihood would be maximized over model parameters. This procedure is possible since we provide a complete statistical description of the system in Eqs. (10)–(12) and the stochastic process is Markovian in the state space of (s_0, s_1, z, y, m). The closely related approach of Suter *et al.* [44] has inspired this idea.

Simple models like ours ignore the spatial localization of mRNA molecules [54, 55] and do not include regulation of gene expression. More realistic modeling presents challenging problems for the future.

It is hoped that the theoretical results presented here will enable a better understanding of stochastic gene expression, both in theory and experimentally.

ACKNOWLEDGMENTS

CSP was supported in part by the Systems Biology Center New York (NIH grant P50GM071558). ARS was supported by NSF grant DMS-1009575 and a fellowship from NYU's GSAS. Arjun Raj and Virzhiniya Lekova of the University of Pennsylvania were helpful in providing insights into the experimental aspects of studying stochastic gene expression. Arjun's advice throughout the course of this project has been greatly appreciated. Thank you to the anonymous referees for their kind remarks.

Appendix A: Derivation of the equations for $u_0(x)$ and $v_0(x)$

To obtain Eqs. (17) and (18), the method characteristics is applied to Eqs. (13)–(16). σ is used as the process variable. $x = x_0 e^{-\sigma}$ are the v characteristics and $x = 1 - (1 - x_0)e^{-\sigma}$ are the u characteristics. The solutions along the respective characteristics are $v = v_0 \tilde{F}_{\mathcal{T}_0} e^{\sigma}$ and $u = u_0 \tilde{F}_{\mathcal{T}_1} e^{\sigma}$. Note that $h_{\mathcal{T}_i} = -\frac{d}{d\sigma} \log \tilde{F}_{\mathcal{T}_i}$. This solution is substituted into Eqs. (15) and (16). Since the solution above the leading characteristic is zero, the upper limits of integration become $\sigma = -\log(1 - x)$ and $\sigma = -\log(x)$ in Eqs. (15) and (16) respectively. Making the change of variables $\sigma = \log\left(\frac{1-x_0}{1-x}\right)$ and $\sigma = \log\left(\frac{x_0}{x}\right)$ in the two equations separately give Eqs. (17) and (18).

The steady-state marginal distribution, $v_{\mathrm{ss}}(x)$, is obtained by integrating over all possible elapsed times, $\sigma = 0, \ldots, -\log(x)$. The solution at a particular elapsed time is related to the solution at zero elapsed time through the change of variables $\sigma = \log\left(\frac{x_0}{x}\right)$, giving

$$\begin{aligned} v_{\mathrm{ss}}(x) &= \int_0^{-\log(x)} v(x, \sigma)\, d\sigma \\ &= \int_x^1 v\left(x, \log\left(\frac{x_0}{x}\right)\right) \frac{dx_0}{x_0} \\ &= \int_x^1 v_0(x_0)\, \tilde{F}_{\mathcal{T}_0}\left(\log\left(\frac{x_0}{x}\right)\right) \frac{dx_0}{x}, \end{aligned} \tag{A1}$$

which is the result given in Eq. (19). Eq. (20) is obtained similarly with the change of variables $\sigma = \log\left(\frac{1-x_0}{1-x}\right)$.

Depending on the dwell time densities, the solutions $v_0(x)$ and $u_0(x)$ can be unbounded, yet integrable, at $x = 1$ and $x = 0$ respectively. It can be shown that $v_0(x) \to 0$ as $x \to 1$ provided $\int_0^{\infty} e^{\sigma} f_{\mathcal{T}_1}(\sigma)\, d\sigma < \infty$ and likewise $u_0(x) \to 0$ as $x \to 0$ provided $\int_0^{\infty} e^{\sigma} f_{\mathcal{T}_0}(\sigma)\, d\sigma < \infty$. When solving this problem numerically, to account for this unbounded behavior of the integrands, the trapezoidal rule is only applied on the interior of the domain. Discretization error near the boundaries prevents the method from being second order accurate when either $v_0(x)$ or $u_0(x)$ is unbounded, but nonetheless the numerical scheme converges. We obtained second order convergence when the solutions are both bounded.

Appendix B: Derivation of the autocovariance formulas

A derivation of Eq. (21) is now presented. The stochastic process for $M(t)$ is assumed stationary. The time shift, τ, in the comparison of $M(t)$ with itself is assumed positive, but will be extended to negative values since $\Phi_{MM}(\tau)$ is necessarily an even function. The number of mRNA molecules at the later time, $M(t+\tau)$, is expressed as a sum of the newly created mRNA molecules since time t, $N(t+\tau)$, and the mRNA molecules that survived from time t to time $t+\tau$, $L(t+\tau)$. By definition, $\Phi_{MM}(\tau) \equiv \mathbb{E}[\, M(t)L(t+\tau)\,] + \mathbb{E}[\, M(t)N(t+\tau)\,] - \mu_M^2$.

The expectation of the products of the random variables $N(t+\tau)$, $L(t+\tau)$ with the random $M(t)$ is evaluated by conditioning on $M(t)$. Once $M(t)$ is given, $\mathbb{E}[L(t+\tau)|M(t)] = M(t)e^{-\delta|\tau|}$ due to the exponential degradation of mRNA, while $N(t+\tau)$ is independent of $M(t)$. The calculation proceeds as

$$\begin{aligned}
&\mathbb{E}[M(t)L(t+\tau)] = \mathbb{E}[\, M^2(t)\,]e^{-\delta\tau} = \mathbb{E}[\, Y^2(t) + \mu_M\,]e^{-\delta\tau} \\
&= \mathbb{E}[\, \mathbb{E}[\, Y^2(t)|Z(\cdot)\,]\,]e^{-\delta\tau} + \mu_M e^{-\delta\tau} \\
&= \mathbb{E}\left[\int_{-\infty}^{t} \nu Z(t_1) e^{-\delta(t-t_1)}\, dt_1 \int_{-\infty}^{t} \nu Z(t_2) e^{-\delta(t-t_2)}\, dt_2\right] e^{-\delta\tau} + \mu_M e^{-\delta\tau} \\
&= \nu^2 \int_{-\infty}^{t}\int_{-\infty}^{t} \left(\Phi_{ZZ}(t_2-t_1) + \mu_Z^2\right) e^{-\delta(2t+\tau-t_1-t_2)} dt_1 dt_2 + \mu_M e^{-\delta\tau} \\
&= \frac{\nu^2}{2\delta}\int_{-\infty}^{\tau} \Phi_{ZZ}(s-\tau) e^{-\delta(2\tau-s)}\, ds + \frac{\nu^2}{2\delta}\int_{\tau}^{\infty} \Phi_{ZZ}(s-\tau) e^{-\delta s}\, ds \\
&\quad + \mu_M^2 e^{-\delta t} + \mu_M e^{-\delta\tau}.
\end{aligned} \tag{B1}$$

The second moment of $M(t)$ is related to the second moment of $Y(t)$ in the first line of Eq. (B1) as a consequence of Eq. (6). The second moment of Y is evaluated by conditioning on the gene state history, $Z(\cdot)$. The double integral is simplified to a single integral by making a change of variables: $s = t_2 - t_1 + \tau$ and $s' = t_2 + t_1$. The resulting region of integration is naturally integrated over in two parts, separated at $s = \tau$. The s' integrals are evaluated to give the last line of Eq. (B1).

The contribution to the autocovariance due to the newly created mRNA since time t is also evaluated by conditioning on the gene-state history, $Z(\cdot)$. Once $Z(\cdot)$ is given, $N(t+\tau)$ is independent of $M(t)$ and the expectations of $M(t)$ and $Y(t)$ are equal, again by Eq. (6). Thus,

$$\begin{aligned}
&\mathbb{E}[M(t)N(t+\tau)] = \mathbb{E}[\, \mathbb{E}[Y(t)N(t+\tau)|Z(\cdot)]\,] \\
&= \mathbb{E}[\, \mathbb{E}[Y(t)|Z(\cdot)]\, \mathbb{E}[\, N(t+\tau)|Z(\cdot)]\,] \\
&= \mathbb{E}\left[\int_{-\infty}^{t} \nu Z(t_1) e^{-\delta(t-t_1)}\, dt_1 \int_{t}^{t+\tau} \nu Z(t_2) e^{-\delta(t+\tau-t_2)}\, dt_2\right] \\
&= \nu^2 \int_{-\infty}^{t}\int_{t}^{t+\tau} e^{-\delta(2t+\tau-t_1-t_2)} \left(\Phi_{ZZ}(t_2-t_1) + \mu_Z^2\right) dt_2\, dt_1 \\
&= \frac{\nu^2}{2\delta}\int_{-\infty}^{0} \left(e^{\delta s} - e^{\delta(s-2\tau)}\right) \Phi_{ZZ}(s-\tau)\, ds \\
&\quad + \frac{\nu^2}{2\delta}\int_{0}^{\tau} \left(e^{-\delta s} - e^{\delta(s-2\tau)}\right) \Phi_{ZZ}(s-\tau)\, ds \\
&\quad + \left(1 - e^{-\delta\tau}\right)\mu_M^2.
\end{aligned} \tag{B2}$$

The integral over the gene-state autocovariance can be simplified by making the same change of variables as before. The region of integration in the forth line of Eq. (B2) is an infinite rectangle, which when rotated (and scaled) is naturally decomposed into two regions, one of which is a finite triangle with vertices at $(s, s') = (0, 2t+\tau)$, $(0, 2t-\tau)$, and $(2t, \tau)$.

When contributions from Eq. (B1) and Eq. (B2) are combined, several terms cancel and others sum to give

$$\Phi_{MM}(\tau) = \mu_M e^{-\delta|\tau|} + \frac{\nu^2}{2\delta}\int_{-\infty}^{\infty} \Phi_{ZZ}(\tau - s) e^{-\delta|s|}\, ds, \tag{B3}$$

where τ is now permitted to be any real value. This is the desired result - Φ_{MM} expressed in terms of Φ_{ZZ}.

In order for Eq. (B3) to be useful, Φ_{ZZ} must be expressed in terms of the gene-state dwell time densities, f_{T_0} and f_{T_1}. A derivation of Eq. (23) will now be presented. The autocovariance of the gene-state variable is computed once the stochastic process is stationary, which means that the autocovariance is a function of only the time delay, τ. Autocovariance is necessarily

an even function and so the derivation proceeds under the assumption that τ is positive.

$$\begin{aligned}\Phi_{ZZ}(\tau) &\equiv \mathbb{E}[\, Z(t)Z(t+\tau)\,] - \mu_Z^2 \\ &= \Pr\{Z(t)=1\}\cdot \Pr\{Z(t+\tau)=1 \,|\, Z(t)=1\} - \mu_Z^2 \\ &= \mu_Z \Pr\{Z(t+\tau)=1 \,|\, Z(t)=1\} - \mu_Z^2 \\ &= \mu_Z \sum_{k=0}^{\infty} \Pr\{Z(t+\tau)=1 \text{ after } k \text{ returns} \,|\, Z(t)=1\} - \mu_Z^2. \end{aligned} \tag{B4}$$

In Eq. (B4), $\Pr\{Z(t+\tau)=1 \,|\, Z(t)=1\}$ is computed by summing the probabilities of all the mutually exclusive ways this event can occur; k counts the number of times the gene returns to the on state after switching to the off state. As long as t is considered arbitrary subject only to the gene being in the on state at time t, the process will be *instantaneously stationary*.

The case in which Z remains in the on state, $k = 0$, is special. In terms of renewal theory [56], this is the stationary probability of the residual life-time exceeding τ:

$$\Pr\{Z(t+\tau)=1 \text{ and } k=0\} = \frac{1}{\mu_{T_1}} \int_{\tau}^{\infty} \tilde{F}_{T_1}(\sigma)\, d\sigma. \tag{B5}$$

For one return to the on state, $k = 1$, a succession of three independent events must occur in $(t, t+\tau)$: the gene must switch off after initial observation at an arbitrary time t in the steady state, which has a probability $\tilde{F}_{T_1}/\mu_{T_1}$; then the gene must switch on, which has a probability density of f_{T_0}; and finally the gene must *not* switch off again, which has a probability $\tilde{F}_{T_1}$. This calculation will proceed in the Fourier domain where convolutions become products. The Fourier transform of the joint probability function for $k = 1$ is

$$\frac{1}{\mu_{T_1}} \frac{1-\hat{f}_{T_1}(\omega)}{i\omega} \hat{f}_{T_0}(\omega) \frac{1-\hat{f}_{T_1}(\omega)}{i\omega}, \tag{B6}$$

where we have used the fact that the Fourier transform of $\tilde{F}_{T_1}(\tau)H(\tau)$ is $\frac{1}{i\omega}\left(1-\hat{f}_{T_1}(\omega)\right)$.

Each additional return to the on state results in another factor of $\hat{f}_{T_0}(\omega)\hat{f}_{T_1}(\omega)$ corresponding to a switch off and then back on. The sum over k is a convergent geometric series as $|\hat{f}_{T_0}(\omega)\hat{f}_{T_1}(\omega)| < 1$ for $\omega \neq 0$. Thus,

$$\Phi_{ZZ}(\tau) = \frac{\mu_Z}{\mu_{T_1}} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{i\omega\tau}}{(i\omega)(i\omega)} \frac{\left(1-\hat{f}_{T_1}(\omega)\right)^2 \hat{f}_{T_0}(\omega)}{1-\hat{f}_{T_0}(\omega)\hat{f}_{T_1}(\omega)}\, d\omega + \int_{\tau}^{\infty} \tilde{F}_{T_1}(\sigma)\, d\sigma \right] - \mu_Z^2. \tag{B7}$$

The Fourier transform of $\int_{\tau}^{\infty} \tilde{F}_{T_1}(\sigma)\, d\sigma H(\tau)$ is $\frac{\mu_{T_1}}{i\omega} - \frac{1}{(i\omega)^2}\left(1-\hat{f}_{T_1}(\omega)\right)$, which can be used to combine the first two terms in Eq. (B7). Eq. (23) is obtained by taking $\frac{1}{i\omega}\frac{1}{i\omega}$ as two time domain integrations and switching the order of those integrations. In practice, it may be simpler to evaluate one or both of the integrations in the frequency domain. Notice that τ is replaced by $|\tau|$ in Eq. (23) to force Φ_{ZZ} to be an even function. If one wishes to evaluate the mRNA autocovariance from Eq. (22), it is worthwhile to note that the above derivation gives the Fourier transform of $\Phi_{ZZ}(\tau)H(\tau)$. The Fourier transform of $\Phi_{ZZ}(\tau)$ is given as $\hat{\Phi}_{ZZ}(\omega) = \mathcal{F}\{\Phi_{ZZ}(\tau)H(\tau)\}(\omega) + \mathcal{F}\{\Phi_{ZZ}(\tau)H(\tau)\}(-\omega)$.

[1] R. J. Johnston and C. Desplan, Annu. Rev. Cell Dev. Biol. **26**, 689 (2010).
[2] A. Becskei, B. Séraphin, and L. Serrano, EMBO J. **20**, 2528 (2001).
[3] F. J. Isaacs, J. Hasty, C. R. Cantor, and J. J. Collins, Proc. Nat. Acad. Sci. USA **100**, 7714 (2003).
[4] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. van Oudenaarden, Nature **427**, 737 (2004).
[5] M. Acar, A. Becskei, and A. van Oudenaarden, Nature **435**, 228 (2005).
[6] L. Weinberger, J. Burnett, J. Toettcher, A. Arkin, and D. Schaffer, Cell **122**, 169 (2005).
[7] H. Maamar, A. Raj, and D. Dubnau, Science **317**, 526 (2007).
[8] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie, Science **322**, 442 (2008).
[9] J. Z. Kelemen, P. Ratna, S. Scherrer, and A. Becskei, PLoS Biol **8**, e1000332 (2010).
[10] C. Zong, L. So, L. A. Sepúlveda, S. O. Skinner, and I. Golding, Mol. Syst. Biol. **6**, 6:440 (2010).
[11] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, PLoS Biol. **4**, e309 (2006).
[12] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, Nat. Rev. Genet. **6**, 451 (2005).
[13] J. M. Raser and E. K. O'Shea, Science **309**, 2010 (2005).
[14] D. R. Larson, R. H. Singer, and D. Zenklusen, Trends Cell Biol. **19**, 630 (2009).
[15] J. Peccoud and B. Ycart, Theor. Popul. Biol. , 222 (1995).

[16] M. Thattai and A. van Oudenaarden, Proc. Nat. Acad. Sci. USA **98**, 8614 (2001).
[17] T. B. Kepler and T. C. Elston, Biophys. J. **81**, 3116 (2001).
[18] J. Paulsson, Nature **427**, 415 (2004).
[19] N. Friedman, L. Cai, and X. S. Xie, Phys. Rev. Lett. **97**, 168302 (2006).
[20] L. Cai, N. Friedman, and X. S. Xie, Nature **440**, 358 (2006).
[21] A. Raj and A. van Oudenaarden, Cell **135**, 216 (2008).
[22] V. Shahrezaei and P. S. Swain, Proc. Nat. Acad. Sci. USA **105**, 17256 (2008).
[23] A. Mugler, A. M. Walczak, and C. H. Wiggins, Phys. Rev. E **80**, 041921 (2009).
[24] A. M. Walczak, A. Mugler, and C. H. Wiggins, Proc. Nat. Acad. Sci. USA **106**, 6529 (2009).
[25] J. M. Pedraza and J. Paulsson, Science **319**, 339 (2008).
[26] B.-L. Xu and Y. Tao, J. Theor. Biol. **243**, 214 (2006).
[27] H. H. McAdams and A. Arkin, Proc. Nat. Acad. Sci. USA **94**, 814 (1997).
[28] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, Nat. Genet. **31**, 69 (2002).
[29] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).
[30] J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer, Curr. Biol. **16**, 1018 (2006).
[31] J. M. Raser and E. K. O'Shea, Science **304**, 1811 (2004).
[32] A. Becskei, B. B. Kaufmann, and A. van Oudenaarden, Nat. Genet. **37**, 937 (2005).
[33] M. Bengtsson, A. Ståhlberg, P. Rorsman, and M. Kubista, Genome Res. **15**, 1388 (2005).
[34] L. Warren, D. Bryder, I. L. Weissman, and S. R. Quake, Proc. Nat. Acad. Sci. USA **103**, 17807 (2006).
[35] D. Zenklusen, D. R. Larson, and R. H. Singer, Nat. Struct. Mol. Biol. **15**, 1263 (2008).
[36] R. Karmakar and I. Bose, Phys. Biol. **1**, 197 (2004).
[37] J. Paulsson, Phys. Life Rev. **2**, 157 (2005).
[38] S. Iyer-Biswas, F. Hayot, and C. Jayaprakash, Phys. Rev. E **79**, 031911 (2009).
[39] L. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding, Nat. Genet. **43**, 554 (2011).
[40] V. Elgart, T. Jia, A. T. Fenley, and R. Kulkarni, Phys. Biol. **8**, 046001 (2011).
[41] R. G. and Roeder, Trends Biochem. Sci. **16**, 402 (1991).
[42] M. Vignali, A. H. Hassan, K. E. Neely, and J. L. Workman, Mol. Cell. Biol. **20**, 1899 (2000).
[43] R. Blossey and H. Schiessel, Biophys. J. **101**, L30 (2011).
[44] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Science **332**, 472 (2011).
[45] A. S. Novozhilov, G. P. Karev, and E. V. Koonin, Briefings Bioinf. **7**, 70 (2006).
[46] S. Iyer-Biswas, Ph.D. thesis, Ohio State University (2009).
[47] M. S. Holla and S. K. Bhattacharya, Ann. Inst. Statist. Math. **17**, 377 (1965).
[48] M. Tang, J. Theor. Biol. **253**, 271 (2008).
[49] R. E. Mortensen, IEEE Trans. Autom. Control **35**, 1245 (1990).
[50] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie, Science **311**, 1600 (2006).
[51] B. Wu, K. D. Piatkevich, T. Lionnet, R. H. Singer, and V. V. Verkhusha, Curr. Opin. Cell Biol. **23**, 310 (2011).
[52] T. Lionnet, K. Czaplinski, X. Darzacq, Y. Shav-Tal, A. L. Wells, J. A. Chao, H. Y. Park, V. de Turris, M. Lopez-Jones, and R. H. Singer, Nat. Methods **8**, 165 (2011).
[53] T. Trcek, D. Larson, A. Moldón, C. Query, and R. Singer, Cell **147**, 1484 (2011).
[54] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, Nat. Methods **5**, 877 (2008).
[55] T. Trcek, J. A. Chao, D. R. Larson, H. Y. Park, D. Zenklusen, S. M. Shenoy, and R. H. Singer, Nat. Protoc. **7**, 408 (2012).
[56] D. R. Cox, *Renewal Theory* (Chapman and Hall, New York, 1962).