Asymptotics of protein number distribution in stochastic gene expression models under burst approximation

Yuntao Lu and Yunxin Zhang*
School of Mathematical Sciences, Fudan University, Shanghai 200433, China

(Received 26 April 2026; accepted 23 September 2026; published 10 November 2026)

The burst approximation is a widely used technique to simplify stochastic gene expression models. However, the dynamics and analytical properties of the protein number distribution in gene expression models under the burst approximation remain largely unexplored. In this study, we propose and systematically analyze surrogate models with multiple gene states and arbitrary burst size distributions. An analytical time-dependent solution to the chemical master equation is derived and then exploited in two directions. First, several fine properties of the protein number distribution are established using functional analysis. For geometrically distributed burst sizes, the distribution is dominated by a scaled negative binomial distribution and is light tailed in certain parameter regimes. Second, we develop efficient algorithms in three settings, enabling fast calculation of the protein number distribution. Furthermore, the approximation error relative to full gene expression models is estimated in terms of low-order moments of the distribution, thereby clarifying the validity of the burst approximation.

DOI: 10.1103/8t2f-bpk1

I. INTRODUCTION

Gene expression is inherently stochastic, as transcription, translation, and molecular degradation are all driven by random molecular events. This intrinsic stochasticity becomes particularly pronounced when relevant biomolecules exist in low copy numbers [1]. Consequently, even genetically identical cells exposed to the same environment may exhibit substantial cell-to-cell variability in messenger RNA (mRNA) and protein levels [2]. Such variability has been directly observed in single-cell and single-molecule experiments, including real-time measurements of transcriptional bursting in bacteria [3] and single-molecule studies of protein production in living cells [4,5]. The underlying molecular mechanism has also been investigated [6–8].

The continuous-time Markov chain is a standard tool for describing interactions among multiple components in a stochastic chemical reaction system, and its Kolmogorov forward equation is conventionally termed the chemical master equation (CME) in such context [9–14]. Gene expression can be readily modeled using the CME by assuming that transcription, translation, and molecular degradation are all first-order reactions [15], as shown in Fig. 1. However, for full gene expression models, the probability distribution of the protein copy number is analytically intractable even when the gene is constitutively active [16]. To overcome this analytical challenge, in Ref. [17], the binomial moment method [18–20] is applied to a model with two gene states (one active, one inactive), where a recurrence relation for binomial moments and analytical expressions for low-order moments are provided. A similar approach is subsequently extended to a model with multiple gene states [21]. Given the difficulty of analyzing full gene expression models, the burst approximation is conceptually proposed as an approximation technique to simplify the full models, building upon the experimental conclusion that mRNAs degrade substantially faster than proteins in many cells [16]. Remarkably, gene expression models under the burst approximation admit fruitful analytical results. In Ref. [16], the generating function method is used to study two models: one in which the gene remains constitutively active and the other in which it switches between an active and an inactive state. For both models, the steady-state probability mass function (PMF) of the protein copy number is derived analytically. More recently, the binomial moment method was applied to models with multiple gene states [22], with a recurrence relation for binomial moments formally obtained and an analytical expression for the PMF presented for models with a single active gene state. Additionally, by establishing a one-to-one correspondence between generic gene expression models and $GI^X/G/\infty$ queueing systems [23,24], non-Markovian models can be studied using queueing theory [25,26]. The correspondence between models under the burst approximation and queueing systems with batch arrivals may explain the substantial simplification introduced by the burst approximation. Furthermore, for certain non-Markovian models, the steady-state distribution of the protein copy number is identical to that of appropriately constructed Markovian counterparts [27,28], thereby allowing the former to be analyzed through the latter.

However, previous studies have several limitations. First, they have mainly focused on the numerical evaluation of the protein number distribution, and few general properties of the distribution are known. Second, previous work has mostly

*Contact author: xyz@fudan.edu.cn

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

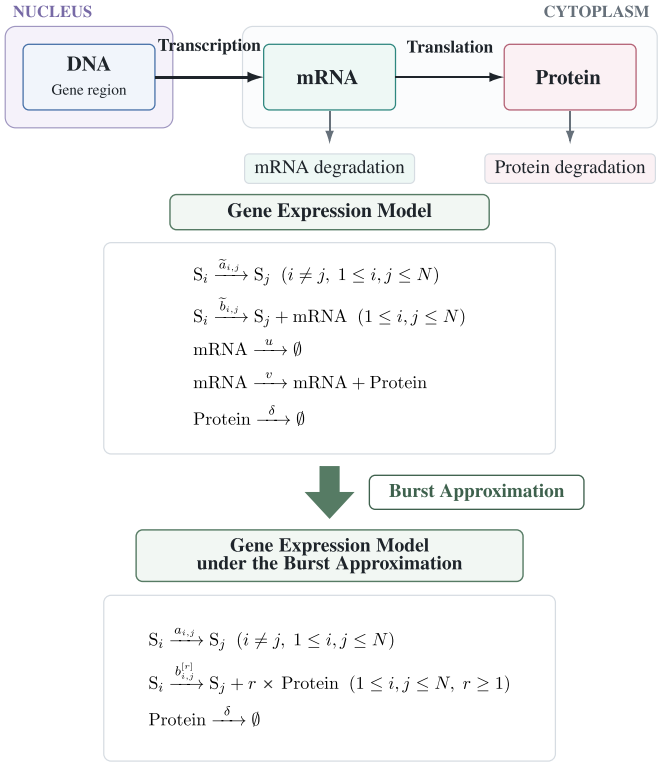


FIG. 1. Workflow of gene expression modeling. In a typical gene expression process, transcription in the nucleus produces mRNAs, which are then translated into proteins in the cytoplasm. Both mRNAs and proteins are subject to degradation. The full gene expression model describes switching among multiple gene states, with productive transitions generating mRNAs, followed by mRNA degradation, translation, and protein degradation. Under the burst approximation, the explicit mRNA intermediate is eliminated and represented by protein bursts.

considered the steady-state distribution, whereas the dynamics has rarely been studied. Moreover, whether the error introduced by the burst approximation is bounded has received limited attention. The validity of the burst approximation has been assessed only for a model where the gene is constitutively active [16].

In this study, we propose and systematically analyze a general gene expression model under the burst approximation, revealing the dynamics and many fine properties of the steady-state distribution of the protein copy number. First, we present an analytical time-dependent solution to the CME, which differs from the formal expression given by the path summation representation [29]. We proceed in two directions. First, the asymptotic behavior of the steady-state distribution is analyzed using functional analysis. Upper bounds for the PMF are established. When the burst size is geometrically distributed, the distribution of the protein copy number is dominated by a scaled negative binomial distribution and is light tailed in certain parameter regimes. An upper bound for the tail probability is also derived. Second, efficient algorithms to calculate the steady-state protein number distribution are developed in three specific settings. Finally, the approximation error introduced by the burst approximation is estimated using functional analysis in terms of low-order moments of the protein number distribution.

II. MODEL DESCRIPTION

We propose a general gene expression model under the burst approximation, represented as a chemical reaction system (1). In this model, the gene has N distinct metastable states, denoted by S_i $(1 \leqslant i \leqslant N)$, and it switches among these states memorylessly. Within each metastable state, proteins are produced in bursts at state-dependent rates, and the gene state may be altered during protein production. Note that a single burst can produce multiple protein molecules simultaneously, with the burst size following a probability distribution. Existing protein molecules degrade at a given rate independent of the gene state and the protein count. Protein autoregulation of gene expression may occur in some biological systems but is not included in model (1) for simplicity [30]. Following the stochastic description of the chemical reaction kinetics [13], model (1) can be mathematically interpreted as a continuous-time Markov chain, whose dynamics is governed by the CME:

$$\begin{aligned} &S_i \xrightarrow{a_{i,j}} S_j \quad (i \neq j,\ 1 \leqslant i, j \leqslant N),\\ &S_i \xrightarrow{b_{i,j}^{[r]}} S_j + r \times \mathrm{Protein} \quad (1 \leqslant i, j \leqslant N,\ r \geqslant 1),\\ &\mathrm{Protein} \xrightarrow{\delta} \emptyset. \end{aligned} \tag{1}$$

Specifically, let $S(t)$, $X(t)$ denote the gene state and the protein count in the system at time t. Let $a_{i,j}$ $(i \neq j,\ 1 \leqslant i, j \leqslant N)$ denote the transition rates between gene states without production of proteins; let $b_{i,j}^{[r]}$ $(1 \leqslant i, j \leqslant N, r \geqslant 1)$ denote the rates of generating r protein molecules in a single burst and the gene transitioning from S_i to S_j; and let δ denote the degradation rate of proteins. Define $a_{i,i} := -\sum_{\substack{k=1\\k\neq i}}^{N} a_{i,k} - \sum_{k=1}^{N}\sum_{r=1}^{\infty} b_{i,k}^{[r]}$ for $1 \leqslant i \leqslant N$, $D_0 := (a_{i,j})_{N\times N}$, $D_r := (b_{i,j}^{[r]})_{N\times N}$ for $r = 1, 2, \ldots$, and $D := D_0 + \sum_{r=1}^{\infty} D_r$. Assume throughout this study that the sequence $\{b_{i,j}^{[r]}\}_{r\geqslant 1}$ has finite moments of all orders, namely, $\sum_{r=1}^{\infty} r^n b_{i,j}^{[r]} < \infty$ for any $n \in \mathbb{N}$ and $1 \leqslant i, j \leqslant N$. In particular, D is well defined. For convenience, we refer to N as the order of the model (1).

III. TIME-DEPENDENT SOLUTION TO CHEMICAL MASTER EQUATION

Let $\mathbb{P}_{i,j}(m;t)$ denote the probability that $S(t) = S_j$ and $X(t) = m$, given the Dirac initial distribution concentrated at $S(0) = S_i$ and $X(0) = 0$. The CME governing $\mathbb{P}_{i,j}(m;t)$ is $(1 \leqslant i, j \leqslant N, m \in \mathbb{N})$

$$\begin{aligned} \frac{\partial}{\partial t}\mathbb{P}_{i,j}(m;t) = &\sum_{s=1}^{N} a_{s,j}\mathbb{P}_{i,s}(m;t) + \sum_{r=1}^{\infty}\sum_{s=1}^{N} b_{s,j}^{[r]}\mathbb{P}_{i,s}(m-r;t)\\ &+ (m+1)\delta\,\mathbb{P}_{i,j}(m+1;t) - m\delta\,\mathbb{P}_{i,j}(m;t), \end{aligned} \tag{2}$$

where $\mathbb{P}_{i,j}(m;t)$ is conventionally taken as zero when $m < 0$. Using the generating function method, we conclude that the

analytical solution to Eq. (2) is, for $m = 0$,

$$\mathbb{P}_{i,j}(0;t) = \boldsymbol{e}_i^\top \mathrm{e}^{D_0 t}\boldsymbol{e}_j + \sum_{k=1}^{\infty}\sum_{l_1,\ldots,l_k\geqslant 1}\int_{\Omega_k}\left[\prod_{s=1}^{k}(1-\alpha(t,t_s))^{l_s}\right] \cdot \boldsymbol{e}_i^\top K(t;t_1,\ldots,t_k;l_1,\ldots,l_k)\boldsymbol{e}_j \mathrm{d}\boldsymbol{t}, \tag{3}$$

and, for $m \geqslant 1$,

$$\mathbb{P}_{i,j}(m;t) = \sum_{k=1}^{\infty}\sum_{l_1,\ldots,l_k\geqslant 1}\int_{\Omega_k}\left[\sum_{\substack{m_1+\cdots+m_k=m\\ 0\leqslant m_1\leqslant l_1,\ldots,0\leqslant m_k\leqslant l_k}}\prod_{s=1}^{k}\binom{l_s}{m_s}\alpha(t,t_s)^{m_s}(1-\alpha(t,t_s))^{l_s-m_s}\right] \cdot \boldsymbol{e}_i^\top K(t;t_1,\ldots,t_k;l_1,\ldots,l_k)\boldsymbol{e}_j \mathrm{d}\boldsymbol{t}. \tag{4}$$

In Eqs. (3) and (4), $\Omega_k := \{(t_1,t_2,\ldots,t_k) \mid 0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_k \leqslant t\} \subseteq \mathbb{R}^k$; $\alpha(t,t_s) := \exp[-\delta(t-t_s)]$; $\binom{l_s}{m_s}$ denotes the binomial coefficient; and $\{\boldsymbol{e}_s\}_{s=1}^N$ are standard basis vectors in $\mathbb{R}^N$ where $\boldsymbol{e}_s$ is the column vector with a 1 in the sth position and 0 elsewhere. The family of functions $\{K(t;t_1,\ldots,t_k;l_1,\ldots,l_k)\}_{k\geqslant 1}$ is defined as follows: $K(t;t_1;l_1) := \mathrm{e}^{D_0 t_1}D_{l_1}\mathrm{e}^{D_0(t-t_1)}$, $K(t;t_1,t_2;l_1,l_2) := \mathrm{e}^{D_0 t_1}D_{l_1}\mathrm{e}^{D_0(t_2-t_1)}D_{l_2}\mathrm{e}^{D_0(t-t_2)}$, and $K(t;t_1,\ldots,t_k;l_1,\ldots,l_k) := \mathrm{e}^{D_0 t_1}D_{l_1}\mathrm{e}^{D_0(t_2-t_1)}D_{l_2}\cdots D_{l_{k-1}}\mathrm{e}^{D_0(t_k-t_{k-1})}D_{l_k}\mathrm{e}^{D_0(t-t_k)}$ for $k \geqslant 3$. The exponential of a square matrix, e.g., C, is defined by $\mathrm{e}^C := \sum_{k=0}^{\infty} C^k/k!$.

Although Eqs. (3) and (4) are not directly suitable for numerical computation, they serve as a starting point for further analysis, which yields fine properties of the steady-state probability distribution. We first proceed by calculating the time-dependent expressions for binomial moments of the protein number distribution and take the temporal limit. This is partially motivated by the well-known binomial moment method. The (matrix-form) binomial moments are defined as $[\mathcal{B}_m(t)]_{i,j} := \sum_{n=m}^{\infty}\binom{n}{m}\mathbb{P}_{i,j}(n;t)$, $1 \leqslant i,j \leqslant N$, $m \in \mathbb{N}$. For technical reasons, we mainly consider the coarse-grained (scalar) binomial moments (in steady state), namely, $B_m := \lim_{t\to\infty}\boldsymbol{\pi}^\top\mathcal{B}_m(t)\mathbf{1}$, $m \in \mathbb{N}$, where $\boldsymbol{\pi} \in \mathbb{R}^{N\times 1}$ is the invariant distribution of the underlying Markov chain characterized by D, and $\mathbf{1} \in \mathbb{R}^{N\times 1}$ is a column vector with all ones. Throughout the study, we assume that D is irreducible, so that it admits a unique invariant distribution. The coarse-grained (scalar) PMF (in steady state) is defined similarly by $P_n := \lim_{t\to\infty}\boldsymbol{\pi}^\top\mathbb{P}(n;t)\mathbf{1}$, $n \in \mathbb{N}$, where $[\mathbb{P}(n;t)]_{i,j} := \mathbb{P}_{i,j}(n;t)$, $1 \leqslant i,j \leqslant N$, $n \in \mathbb{N}$.

We conclude that the binomial moments in steady state are all finite and have the following expressions:

$$B_1 = \frac{1}{\delta}\boldsymbol{\pi}^\top C_1\mathbf{1},$$

$$B_m = \frac{1}{m\delta}\sum_{k=1}^{m}\sum_{\substack{l_1+\cdots+l_k=m\\ l_1,\ldots,l_k\geqslant 1}}\boldsymbol{\pi}^\top C_{l_1}(l_1\delta\boldsymbol{I}_N - D)^{-1} \times C_{l_2}[(l_1+l_2)\delta\boldsymbol{I}_N - D]^{-1}C_{l_3} \cdots [(l_1+\cdots+l_{k-1})\delta\boldsymbol{I}_N - D]^{-1}C_{l_k}\mathbf{1}, \quad m \geqslant 2, \tag{5}$$

where $C_r := \sum_{n=r}^{\infty}\binom{n}{r}D_n$ and $\boldsymbol{I}_N$ denotes the N-dimensional identity matrix. Note that $\{C_r\}_{r\geqslant 1}$ are all well defined since we assume moments of all orders of $\{b_{i,j}^{[r]}\}_{r\geqslant 1}$ exist. In Eq. (5), the summation is taken over all ordered partitions of m into positive integers. For example, when $m = 3$, the summation includes four ordered partitions: $l_1 = 3$; $l_1 = 1, l_2 = 2$; $l_1 = 2, l_2 = 1$; and $l_1 = 1, l_2 = 1, l_3 = 1$. From Eq. (5), one can readily calculate low-order binomial moments and, consequently, important summary statistics such as the noise and the Fano factor. The analytical results are reported in Supplemental Material [31] and are consistent with existing studies [22]. In general, direct computation based on Eq. (5) quickly becomes impractical as m grows, because of the combinatorial enumeration needed to determine the integer partition of m. Actually, matrix computations within the summation are relatively fast; therefore, the computational bottleneck is independent of the order of the model (N) and is combinatorial. We later provide a combinatorial approach to overcome this issue in certain cases.

If $D_r = c_r D_1$ ($r \geqslant 2$) for some constants $c_r \geqslant 0$ ($r \geqslant 2$), and D_1 is diagonal, the model (1) reduces to the one in Ref. [22].

IV. ASYMPTOTIC ANALYSIS OF STEADY-STATE PROBABILITY MASS FUNCTION

We derive the upper bound for the PMF by repeatedly applying an analog to Theorem 4.1.2 in Ref. [32], the Hölder inequality, and the submultiplicativity of operator norm. Specifically, we have (for $n \geqslant 1$)

$$P_n \leqslant \sum_{k=1}^{n}\sum_{\substack{l_1+\cdots+l_k=n\\ l_1,\ldots,l_k\geqslant 1}}\frac{\prod_{i=1}^{k}\|C_{l_i}\|_\infty}{\delta^k l_1(l_1+l_2)\cdots(l_1+\cdots+l_k)}, \tag{6}$$

where the infinity norm of a matrix, denoted by $\|\cdot\|_\infty$, is the maximum absolute row sum of this matrix. By definition, $\|C_r\|_\infty$ has the same units as $b_{i,j}^{[r]}$ ($1 \leqslant i,j \leqslant N$) for $r \geqslant 1$, and the upper bound in Eq. (6) is dimensionless.

In general, Eq. (6) cannot be further simplified. However, when $\{D_r\}_{r\geqslant 1}$ follows a geometric distribution, we obtain elegant results. We note that this assumption is reasonable, since the number of proteins generated during one burst has been experimentally shown to approximately follow a geometric distribution [4,5]. Specifically, we now assume $D_r = \lambda^{r-1}D_1$ ($r \geqslant 1$), where $\lambda \in (0,1)$ is the parameter of a geometric distribution. Equation (6) then reduces to

$$P_n \leqslant \frac{1}{n!}\left(\frac{\lambda}{1-\lambda}\right)^n\left[\frac{\|D_1\|_\infty}{\delta\lambda(1-\lambda)}\right]^{(n)}, \quad n \in \mathbb{N}. \tag{7}$$

For any $\zeta > 0$ and $m \in \mathbb{N}$, $\zeta^{(m)} := \Gamma(\zeta+m)/\Gamma(\zeta)$ is the Pochhammer symbol, where $\Gamma(\zeta)$ is the Gamma function. A combinatorial identity is developed to prove Eq. (7), which overcomes the complexity bottleneck of enumerating ordered partitions of integers. The upper bound in Eq. (7) converges to zero if and only if $\lambda \in (0,1/2)$ or $\lambda = 1/2$ and $\|D_1\|_\infty/\delta < 1/4$. In particular, when $\lambda \in (0,1/2)$, it follows from Eq. (7) that the PMF of the protein copy number is bounded from above by a constant multiple of a negative binomial

distribution. See Fig. S1 of Supplemental Material [31] for numerical verification.

When $\lambda \in (0, 1/2)$, the steady-state probability distribution of the protein copy number is light tailed, since the moment generating function $\mathcal{M}[\{P_n\}](s) := \sum_{n=0}^{\infty} P_n \exp(ns)$ is finite for $s \in (0, \ln(1-\lambda) - \ln\lambda)$. Based on the Chernoff method [33], we also estimate the tail probability of the protein copy number. Specifically, for $m > \|D_1\|_\infty/[\delta(1-2\lambda)(1-\lambda)]$, we have

$$\sum_{n=m}^{\infty} P_n \leqslant \left[\frac{\|D_1\|_\infty + m\delta\lambda(1-\lambda)}{m\delta(1-\lambda)^2}\right]^m \times \left[1 + \frac{m\delta\lambda(1-\lambda)}{\|D_1\|_\infty}\right]^{\frac{\|D_1\|_\infty}{\delta\lambda(1-\lambda)}}. \tag{8}$$

The above estimation of the moment generating function and the tail probability generalizes the heavy-tailed law for transcription models [34–36].

V. FAST SOLVER FOR STEADY-STATE PROBABILITY DISTRIBUTION

There are two motivations for examining the recurrence relation among subsequent binomial moments. First, according to Eq. (5), many terms in the summation are repeatedly recomputed if all binomial moments are computed independently. Second, the binomial moment equations [18] are decoupled for the first-order chemical reaction system (1); therefore, higher-order binomial moments are linear combinations of lower-order binomial moments in steady state. In this study, we derive the recurrence relation directly from Eq. (5). A formal but more direct approach is to apply the binomial moment method as performed in Ref. [22]. Based on Eq. (5), we have

$$\begin{aligned} &\mathcal{B}_0 := \boldsymbol{I}_N, \quad \mathcal{B}_m := \left(\sum_{n=1}^{m} \mathcal{B}_{m-n} C_n\right)(m\delta \boldsymbol{I}_N - D)^{-1}, \\ &B_m = \frac{1}{m\delta}\boldsymbol{\pi}^\top \left(\sum_{n=1}^{m} \mathcal{B}_{m-n} C_n\right)\mathbf{1}, \quad m \geqslant 1. \end{aligned} \tag{9}$$

A numerical implementation of Eq. (9) is given in Algorithm S1 of Supplemental Material [31].

According to Algorithm S1 of Supplemental Material [31], binomial moments up to arbitrary orders can be obtained recursively. Formally, the PMF can then be constructed based on the following identity [20]:

$$P_n = \sum_{m=n}^{\infty} (-1)^{m-n} \binom{m}{n} B_m, \quad n \in \mathbb{N}. \tag{10}$$

However, for the series (10) to converge, the term $\binom{m}{n} B_m$ should converge to zero as m grows, implying that the sequence $\{B_m\}_{m\in\mathbb{N}}$ converges to zero. We prove that when the burst size follows a geometric distribution with parameter $\lambda \in (0, 1/2)$, the PMF can be reconstructed through Eq. (10). When the burst size follows a geometric distribution with parameter $\lambda \in [1/2, 1)$, the series (10) may diverge and the PMF cannot be readily recovered. In Fig. S2 of Supplemental Material [31], we provide examples where all orders of binomial moments exist but $\lim_{m\to\infty} B_m = \infty$.

Additionally, the truncation error may be introduced during the computation of $\{C_r\}_{r\geqslant 1}$. Fortunately, analytical expressions for $\{C_r\}_{r\geqslant 1}$ can be derived under three distinct yet reasonable assumptions.

(1) *Burst size follows geometric distribution*. This assumption has been introduced above. When $D_r = \lambda^{r-1} D_1$ $(r \geqslant 1)$ for some given $\lambda \in (0, 1)$, we have (for $r \geqslant 1$)

$$C_r = \frac{\lambda^{r-1}}{(1-\lambda)^{r+1}} D_1, \quad D = D_0 + \frac{1}{1-\lambda} D_1. \tag{11}$$

Across several examples, we compare our results with those obtained using the stochastic simulation algorithm [37–39] and the finite state projection algorithm [40], as shown in Fig. 2.

(2) *Burst size follows Poisson distribution*. When the burst size follows a Poisson distribution, namely, $D_r = (\alpha^{r-1}/r!)D_1$, $r \geqslant 1$ for some positive real number α, we also obtain analytical expressions (for $r \geqslant 1$)

$$C_r = \frac{\alpha^{r-1}}{r!} \mathrm{e}^{\alpha} D_1, \quad D = D_0 + \frac{1}{\alpha}(\mathrm{e}^{\alpha} - 1) D_1. \tag{12}$$

(3) *Burst becomes negligible for large sizes*. During the gene expression process, the probability that large numbers of proteins are simultaneously produced is negligible. Therefore, we may assume $D_r = \mathbf{0}_{N\times N}$ for all r strictly exceeding a given threshold $r_{\max}$. In such cases, $C_r = \mathbf{0}_{N\times N}$ for $r \geqslant r_{\max} + 1$, and the sum $C_r = \sum_{n=r}^{r_{\max}} \binom{n}{r} D_n$, $1 \leqslant r \leqslant r_{\max}$, involves only finitely many terms. This assumption is also a prerequisite for numerically studying models in Eq. (2) via the stochastic simulation algorithm and the finite state projection algorithm, since they only admit a finite number of reaction pathways. Empirically, as the truncation level $r_{\max}$ increases, the computed probability distribution converges rapidly to the exact distribution corresponding to the full, untruncated model. See Fig. 3 for examples.

VI. SPECIAL CASE: MODEL OF ORDER 1

We now connect our results to those in Ref. [16] by studying the case where $N = 1$ in Eq. (1) and the burst size of proteins is geometrically distributed. This setting corresponds to a constitutively active gene with translation events occurring according to a Poisson process. Assume that $\{D_r\}_{r\geqslant 1}$ follows a geometric distribution with parameter $\lambda \in (0, 1/2)$ and $N = 1$. Note that D_1 is a non-negative real number. Based on Eqs. (5) and (10), we derive that the steady-state distribution of the protein count is a negative binomial distribution (for $n \in \mathbb{N}$):

$$P_n = \frac{\lambda^n}{n!}\left[\frac{D_1}{\delta\lambda(1-\lambda)}\right]^{(n)} (1-\lambda)^{D_1/[\delta\lambda(1-\lambda)]}, \tag{13}$$

which is consistent with existing results [16].

VII. ANALYSIS OF APPROXIMATION ERROR

Finally, we compare low-order moments of the protein number distribution in full gene expression models and

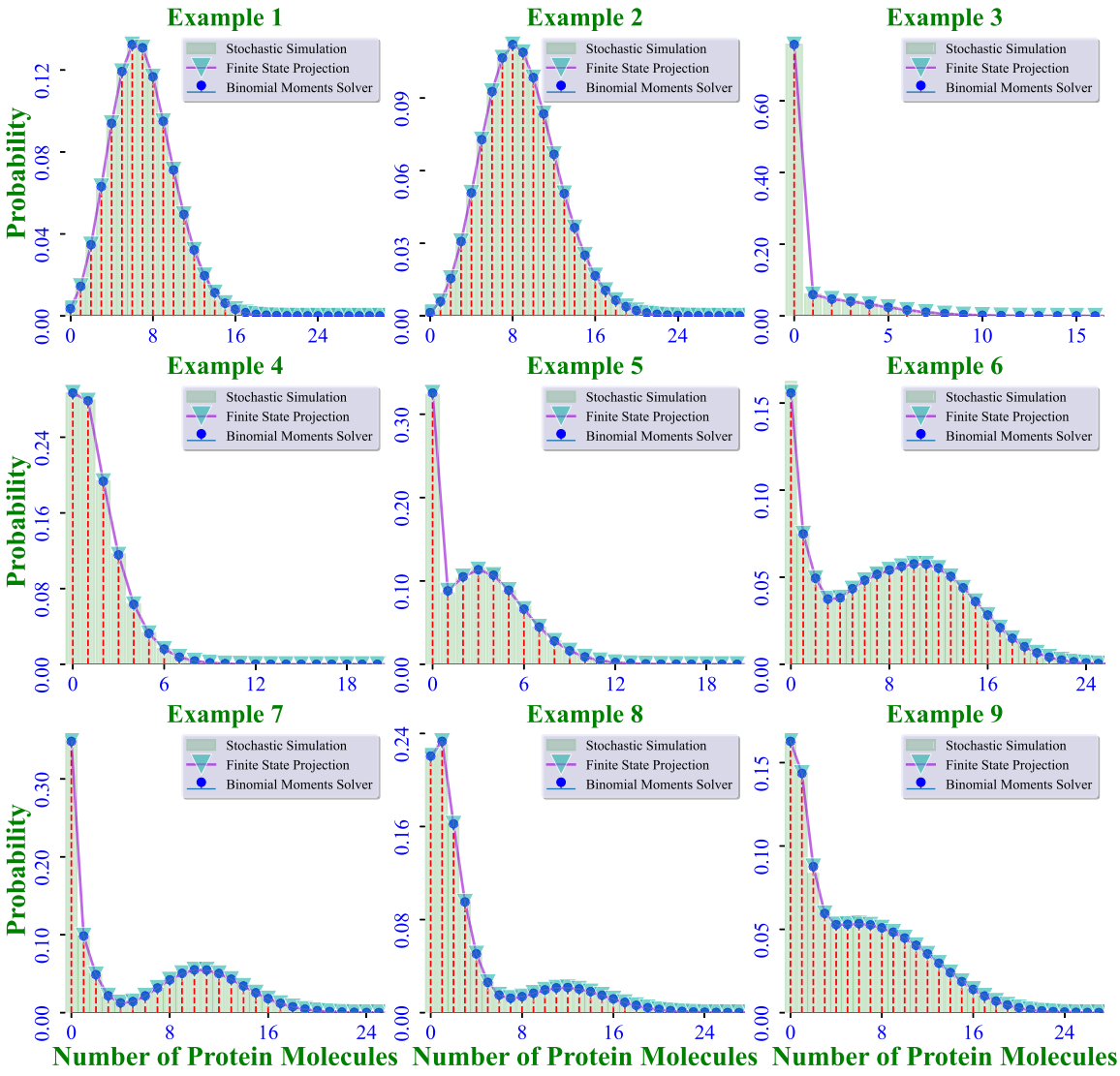


FIG. 2. Probability distribution of protein copy number verified by stochastic simulation and finitestate projection. This figure presents the steady-state probability distributions of the protein copy number for nine representative examples. For each case, our approach (abbreviated *Binomial Moments Solver*) is illustrated by stem plots, the stochastic simulation results are shown as semi-transparent histograms, and the finite state projection results are displayed as line plots with markers. The PYTHON package `GillesPy2` [41] is used with C++ solver to implement the stochastic simulation algorithm. Model parameters and computational details, including the truncation strategy, are described in Supplemental Material [31].

models under the burst approximation. Upper bounds of the difference between low-order moments are established using functional analysis. Recall that, as shown in Fig. 1, the burst approximation is conceptually an approximation technique to simplify the following full gene expression model:

$$\begin{aligned} &S_i \xrightarrow{\widetilde{a}_{i,j}} S_j \quad (i \neq j, \quad 1 \leqslant i, j \leqslant N), \\ &S_i \xrightarrow{\widetilde{b}_{i,j}} S_j + \text{mRNA} \quad (1 \leqslant i, j \leqslant N), \\ &\text{mRNA} \xrightarrow{u} \emptyset, \\ &\text{mRNA} \xrightarrow{v} \text{mRNA} + \text{Protein}, \\ &\text{Protein} \xrightarrow{\delta} \emptyset. \end{aligned} \tag{14}$$

In Eq. (14), the parameter δ carries the same meaning as in Eq. (1), while $\widetilde{a}_{i,j}$, $\widetilde{b}_{i,j}$ $(1 \leqslant i, j \leqslant N)$, v, and u denote the transition rates among gene states without transcription, the transition rates with production of one mRNA molecule, the translation rate, and the mRNA degradation rate, respectively.

Under the notations in Eqs. (1) and (14), two models are related by the following identities:

$$\begin{aligned} b_{i,j}^{[r]} &= \left(\frac{v}{u+v}\right)^r \frac{u}{u+v}\,\widetilde{b}_{i,j}, \quad 1 \leqslant i, j \leqslant N, \ r \geqslant 1, \\ a_{i,j} &= \widetilde{a}_{i,j} + \frac{u}{u+v}\widetilde{b}_{i,j}, \quad 1 \leqslant i \neq j \leqslant N. \end{aligned} \tag{15}$$

The relation (15) can be interpreted as follows. Once transcribed, an mRNA molecule is subject to two competing

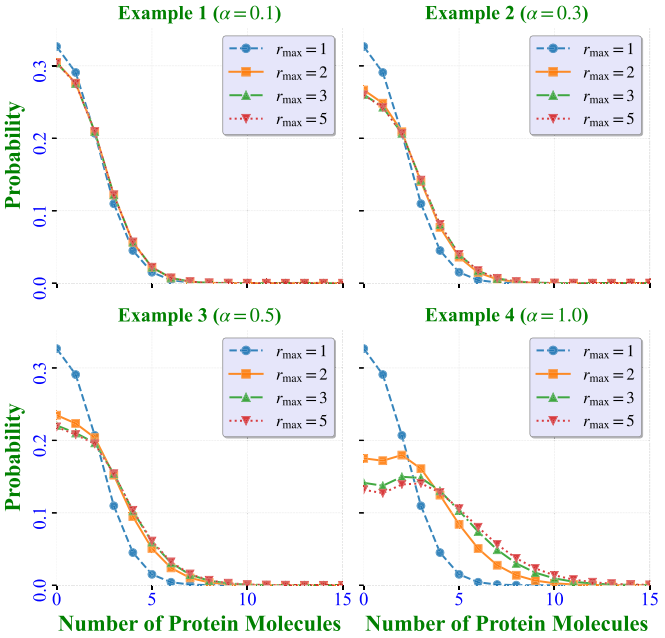


FIG. 3. Probability distribution of protein copy number with different truncation levels. The burst size is assumed to follow a Poisson distribution parametrized by α. The figure shows the steady-state PMF of the protein number with varying truncation levels r_{max}. The four panels correspond to $\alpha = 0.1$, 0.3, 0.5, and 1.0, and within each panel the distributions obtained with $r_{max} = 1$, 2, 3, and 5 are compared. The off-diagonal entries of D_0 are zero, $D_1 = ((0.2, 0.2, 0.1, 0.8), (0.2, 0.2, 0.1, 0.8), (0.1, 0.1, 0.1, 0), (0.8, 0.8, 0, 0.8))$, and $\delta = 1$.

reaction pathways, namely, degradation and translation. More specifically, this can be seen as the competing binding of decay complexes that promote degradation and the recruitment of initiation factors that engage the ribosome for translation. Since the probability of initiating translation rather than degradation is $v/(u+v)$, the number of protein molecules produced from a single mRNA molecule follows a geometric distribution with parameter $u/(u+v)$. Hence, Eq. (15) readily follows.

Low-order moments of mRNA and protein copy numbers in Eq. (14) can be formally obtained by applying the binomial moment method [17,21]. By comparison with Eq. (5), we conclude that the first-order binomial moment, namely, the expectation of the protein copy number, remains exact under the burst approximation, while higher-order binomial moments are explicitly altered when the burst approximation is applied. In particular, upper bounds can be derived for the absolute difference between the corresponding low-order binomial moments of the two models. For example, denoting the second-order binomial moment of the protein copy number in Eq. (14) by $\widetilde{B}_2$, we have

$$|B_2 - \widetilde{B}_2| \leqslant \frac{v(u+v)^2}{2u^3(u+\delta)}\|D_1\|_\infty + \frac{(u+v)^4}{2u^5\delta(u+\delta)}\|D_1\|_\infty^2\|D\|_\infty. \quad (16)$$

By Eq. (16), the absolute difference converges to zero at the rate of $O(u^{-2})$ when $u \to \infty$ and the other parameters remain

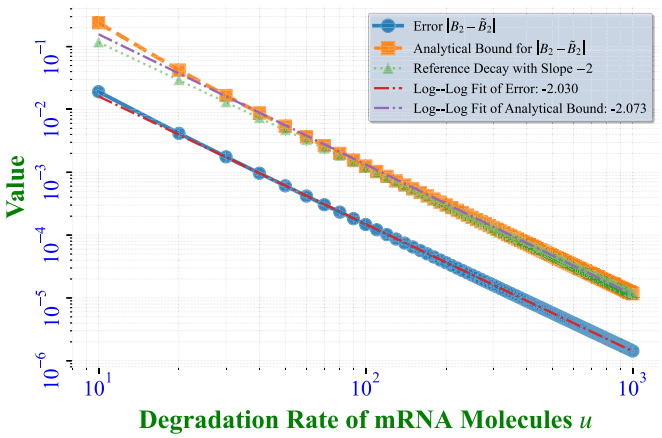


FIG. 4. Approximation error with respect to second-order binomial moments. The absolute error $|B_2 - \widetilde{B}_2|$ and its analytical upper bound (16) are plotted against the mRNA degradation rate u on a logarithmic scale. Dashed lines with square markers indicate the bound, solid lines with circular markers indicate the computed absolute error, and the dotted line represents a reference decay rate. Dash-dotted lines show the corresponding log-log fitted slopes. In this example, D_0, D_1, and δ are the same as those in Fig. 3. Additionally, $v = 2$ and $u \in [10, 1000]$.

fixed, as numerically verified in Fig. 4. Notably, $u/\delta \gg 1$ does not guarantee the validity of the burst approximation.

VIII. CONCLUSIONS

In this study, we systematically analyze gene expression models under the burst approximation with multiple gene states and arbitrary burst size distributions. An analytical solution to the corresponding CME is derived and exploited in two directions. On the one hand, we derive several inequalities for the PMF, the moment generating function, and the tail probability, using functional analysis. When the burst size is geometrically distributed with its parameter in $(0, 1/2)$, the probability distribution of the protein copy number is bounded by a scaled negative binomial distribution and is light tailed. The tail probability is estimated using the Chernoff method. On the other hand, we propose fast solvers for the steady-state distribution of the protein count in three special cases, namely, when the burst size is geometrically distributed, Poisson distributed, or negligible beyond a certain threshold. Additionally, we show that our results are consistent with those in Ref. [16] and examine the validity of the burst approximation by estimating the approximation error in terms of low-order binomial moments. We conclude that, although Eq. (1) is usually seen as a surrogate for full gene expression models, only the expectation of the protein number distribution is preserved, whereas higher-order moments generally differ.

Several directions for future work naturally arise from this study. First, the analytical solution to the CME may be further analyzed to estimate the mixing time of the corresponding Markov chain. Second, when the burst size is geometrically distributed with parameter $\lambda \in [1/2, 1)$, we have identified examples in which the binomial moment method fails to recover

the PMF. The computation of the PMF in such cases remains open, and general conditions guaranteeing the success of the binomial moment method also remain to be established. Third, models formulated via the CME implicitly assume the Markov property, whereas many physical systems are highly non-Markovian. This suggests that stochastic gene expression modeling should be extended to more general stochastic processes, such as those described by the generalized chemical master equation [42]. Finally, in practice, the model validity is typically assessed from a top-down perspective. For example, the model parameters can be estimated by fitting the predicted protein number distribution to experimental data [12,43,44]. The development of appropriate statistical inference frameworks remains an important research objective.

ACKNOWLEDGMENTS

We thank Ruiqi Gao and Pinchen Xie for fruitful discussions. Y.L. was supported by National Natural Science Foundation of China (125B10002) and FDUROP (Fudan Undergraduate Research Opportunities Program) (24260). Y.Z. was supported by National Key R&D Program of China (2024YFA1012401), the Science and Technology Commission of Shanghai Municipality (23JC1400501), and National Natural Science Foundation of China (12241103).

DATA AVAILABILITY

The data that support the findings of this article are openly available [45].

[1] G. W. Li and X. S. Xie, Central dogma at the single-molecule level in living cells, Nature (London) **475**, 308 (2011).

[2] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, Stochastic mRNA synthesis in mammalian cells, PLoS Biol. **4**, e309 (2006).

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

[4] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie, Probing gene expression in live cells, one protein molecule at a time, Science **311**, 1600 (2006).

[5] L. Cai, N. Friedman, and X. S. Xie, Stochastic protein expression in individual cells at the single molecule level, Nature (London) **440**, 358 (2006).

[6] A. Klindziuk and A. B. Kolomeisky, Theoretical investigation of transcriptional bursting: A multistate approach, J. Phys. Chem. B **122**, 11969 (2018).

[7] A. Klindziuk, B. Meadowcroft, and A. B. Kolomeisky, A mechanochemical model of transcriptional bursting, Biophys. J. **118**, 1213 (2020).

[8] A. Klindziuk and A. B. Kolomeisky, Understanding the molecular mechanisms of transcriptional bursting, Phys. Chem. Chem. Phys. **23**, 21399 (2021).

[9] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed. (Elsevier, Amsterdam, 2007).

[10] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences*, 4th ed. (Springer, Berlin, 2009).

[11] D. F. Anderson and T. G. Kurtz, *Stochastic Analysis of Biochemical Systems* (Springer, Cham, 2015).

[12] D. Schnoerr, G. Sanguinetti, and R. Grima, Approximation and inference methods for stochastic biochemical kinetics—A tutorial review, J. Phys. A: Math. Theor. **50**, 093001 (2017).

[13] W. E, T. Li, and E. Vanden-Eijnden, *Applied Stochastic Analysis* (American Mathematical Society, Providence, RI, 2019).

[14] H. Qian and H. Ge, *Stochastic Chemical Reaction Systems in Biology*, Lecture Notes on Mathematical Modelling in the Life Sciences (Springer, Cham, 2021).

[15] J. Chen, Single-cell stochastic gene expression dynamics: Mathematical models and analytical theory, Sci. Sin. Math. **56**, 1073 (2026).

[16] V. Shahrezaei and P. S. Swain, Analytical distributions for stochastic gene expression, Proc. Natl. Acad. Sci. USA **105**, 17256 (2008).

[17] H. H. Qiu and T. S. Zhou, Exact distributions in the full models of gene expression, Sci. Sin. Math. **47**, 1863 (2017).

[18] B. Barzel and O. Biham, Binomial moment equations for stochastic reaction systems, Phys. Rev. Lett. **106**, 150602 (2011).

[19] B. Barzel and O. Biham, Stochastic analysis of complex reaction networks using binomial moment equations, Phys. Rev. E **86**, 031126 (2012).

[20] J. Zhang, Q. Nie, and T. Zhou, A moment-convergence method for stochastic analysis of biochemical reaction networks, J. Chem. Phys. **144**, 194109 (2016).

[21] H. Qiu, B. Zhang, and T. Zhou, Influence of complex promoter structure on gene expression, J. Syst. Sci. Complex. **32**, 600 (2019).

[22] M. Chen, S. Luo, M. Cao, C. Guo, T. Zhou, and J. Zhang, Exact distributions for stochastic gene expression models with arbitrary promoter architecture and translational bursting, Phys. Rev. E **105**, 014405 (2022).

[23] L. Liu, B. R. K. Kashyap, and J. G. C. Templeton, On the $GI^X/G/\infty$ system, J. Appl. Probab. **27**, 671 (1990).

[24] J. Szavits-Nossan and R. Grima, Solving stochastic gene-expression models using queueing theory: A tutorial review, Biophys. J. **123**, 1034 (2024).

[25] T. Jia and R. V. Kulkarni, Intrinsic noise in stochastic models of gene expression with molecular memory and bursting, Phys. Rev. Lett. **106**, 058102 (2011).

[26] N. Kumar, A. Singh, and R. V. Kulkarni, Transcriptional bursting in gene expression: Analytical results for general stochastic models, PLoS Comput. Biol. **11**, e1004292 (2015).

[27] J. Zhang and T. Zhou, Markovian approaches to modeling intracellular reaction processes with molecular memory, Proc. Natl. Acad. Sci. USA **116**, 23542 (2019).

[28] Z. Wang, Z. Zhang, and T. Zhou, Analytical results for non-Markovian models of bursty gene expression, Phys. Rev. E **101**, 052406 (2020).

[29] S. X. Sun, Path summation formulation of the master equation, Phys. Rev. Lett. **96**, 210602 (2006).

[30] R. Grima, D. R. Schmidt, and T. J. Newman, Steady-state fluctuations of a genetic feedback loop: An exact solution, J. Chem. Phys. **137**, 035104 (2012).

[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/8t2f-bpk1 for detailed mathematical derivations, algorithmic procedures, numerical validations, and additional figures supporting the results reported in the main text.

[32] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (Johns Hopkins University Press, Baltimore, MD, 2013).

[33] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press, Oxford, 2013).

[34] L. Ham, R. D. Brackston, and M. P. H. Stumpf, Extrinsic noise and heavy-tailed laws in gene expression, Phys. Rev. Lett. **124**, 108101 (2020).

[35] L. Ham, D. Schnoerr, R. D. Brackston, and M. P. H. Stumpf, Exactly solvable models of stochastic gene expression, J. Chem. Phys. **152**, 144106 (2020).

[36] Y. Lu and Y. Zhang, Stochastic kinetics of mRNA molecules in a general transcription model, Biophys. J. **124**, 4102 (2025).

[37] D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. **22**, 403 (1976).

[38] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. **81**, 2340 (1977).

[39] D. T. Gillespie, Stochastic simulation of chemical kinetics, Annu. Rev. Phys. Chem. **58**, 35 (2007).

[40] B. Munsky and M. Khammash, The finite state projection algorithm for the solution of the chemical master equation, J. Chem. Phys. **124**, 044104 (2006).

[41] S. Matthew, F. Carter, J. Cooper, M. Dippel, E. Green, S. Hodges, M. Kidwell, D. Nickerson, B. Rumsey, J. Reeve, L. R. Petzold, K. R. Sanft, and B. Drawert, GillesPy2: A biochemical modeling framework for simulation driven biological discovery, Lett. Biomath. **10**, 87 (2023).

[42] T. Aquino and M. Dentz, Chemical continuous time random walks, Phys. Rev. Lett. **119**, 230601 (2017).

[43] J. Kim and J. C. Marioni, Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data, Genome Biol. **14**, R7 (2013).

[44] S. Luo, Z. Zhang, Z. Wang, X. Yang, X. Chen, T. Zhou, and J. Zhang, Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model, R. Soc. Open Sci. **10**, 221057 (2023).

[45] Y. Lu and Y. Zhang, GitHub repository, 2026, https://github.com/yuntao2022/Stochastic_Translation.