Mean-field theory accurately captures the variation of copy number distributions across the mRNA's life cycle

Juraj Szavits-Nossan* and Ramon Grima†
School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, United Kingdom

(Dated: August 24, 2021)

We consider a stochastic model where a gene switches between two states, an mRNA transcript is released in the active state and subsequently it undergoes an arbitrary number of sequential unimolecular steps before being degraded. The reactions effectively describe various stages of the mRNA life cycle such as initiation, elongation, termination, splicing, export and degradation. We construct a novel mean-field approach that leads to closed-form steady-state distributions for the number of transcript molecules at each stage of the mRNA life cycle. By comparison with stochastic simulations, we show that the approximation is highly accurate over all of parameter space, independent of the type of expression (constitutive or bursty) and of the shape of the distribution (unimodal, bimodal and nearly bimodal). The theory predicts that in a population of identical cells, any bimodality is gradually washed away as the mRNA progresses through its life cycle.

I. INTRODUCTION

In recent years, the study of stochastic models of gene expression has received wide attention [1]. In the absence of transcriptional feedback, these models are often composed of first-order reactions, in which case exact expressions can be derived from the master equation for all the moments of the distribution of mRNA numbers. This is because for linear propensities, the moment equations are closed and can be solved straightforwardly [2].

However the exact closed-form solution of the master equation for the probability distribution of mRNA numbers is a much harder problem. Popular methods such as the Poisson representation of Gardiner [3] and those devised more recently by Jahnke and Huisanga [4] are not applicable to models of gene expression composed of only first-order reactions because of the presence of a reaction of the type $G \to G + M$ which models transcription of mRNA (M) from the active state of a promoter (G). Hence the general solution of these models is still an open research question. However progress has been achieved on the solution of specific models.

A two-state model (commonly known as the telegraph model) whereby a gene switches between two states (an active and an inactive state) and mRNA is transcribed in the active state, has been solved exactly in steady-state [5] and in time [6] for the marginal distribution of mRNA numbers. Various other extensions of this model to include more biological realism have also been solved exactly or approximately for the marginal distribution of mRNA numbers. These include models that account for more than two gene states [7–9], for polymerase dynamics [10], for leaky expression from the inactive state [11], cell-to-cell variability (static and dynamic) [11, 12], replication and binomial partitioning due to cell division [13], cell cycle duration variability [14] and modulation under environmental changes [15].

We here consider a different type of extension of the telegraph model that has recently received attention in three different biological contexts: (i) RNA polymerase (RNAP) movement along a gene during transcription [16]; (ii) multi-step splicing [17]; (iii) nuclear retention of mRNA [18]. In this stochastic model, a gene switches between two states (an active state G and an inactive one G^*), produces a transcript in the active state and subsequently it undergoes an arbitrary number L of sequential unimolecular steps, leading to L forms of the transcript (denoted by P_i, where $i \in [1, L]$). Each transcript either is removed after the L steps or else can also be degraded prematurely at an earlier step. A reaction scheme for this model is as follows:

$$\begin{aligned} &G \rightleftharpoons G^*, \quad G \to G + P_1, \\ &P_1 \to P_2 \to ... \to P_L \to \emptyset, \\ &P_1 \to \emptyset, P_2 \to \emptyset, ..., P_L \to \emptyset.. \end{aligned} \tag{1}$$

The L downstream processing steps can be interpreted in various ways. At the coarsest scale, the model with $L = 2$ steps can capture nuclear mRNA and cytoplasmic mRNA dynamics. At a less coarser scale, the model can capture the dynamics of the whole mRNA life cycle from birth (initiation) to death (in the cytoplasm). For example one can associate steps $i = 1, ..., L_1 - 1$ with elongation of the nascent transcript, $i = L_1$ with termination, $i = L_1 + 1, ..., L_2 - 1$ with splicing, $i = L_2$ with export from the nucleus to the cytoplasm and $i = L_2 + 1, ..., L$ with several reaction steps leading to mRNA degradation. For the latter process, for example, the sequence of steps can model the fact that the polyadenylate tails of eukaryotic transcripts are sequentially chewed up before the protein-coding part of the message is degraded [19]. At a fine scale, the model could capture the dynamics of nascent mRNA only, where the L steps represent the unidirectional movement of RNAP (with a nascent mRNA tail attached to it) along the gene. Hence the reaction scheme (1) has myriad molecular interpretations according to the desired application.

* Juraj.Szavits.Nossan@ed.ac.uk
† Ramon.Grima@ed.ac.uk

An exact, steady-state and closed-form solution for the marginal distribution of the number of molecules of P_i is unknown. In this paper, we report an approximate solution to this open problem. Because the method we devise is non-perturbative and does not make assumptions on the strength of correlations between the forms of the product, we find by comparison with stochastic simulations that the approximation is highly accurately across all of parameter space.

The paper is divided as follows. In Section II, we describe the model in detail, formulate its master equation and discuss some known exactly solvable cases. In Section III, we derive exact expressions for the moments of P_i using a lattice path method. In Section IV, we use the first and second moments and mean-field approximation to construct marginal distributions for P_i. We show that while the use of standard mean-field approximation is valid in only certain regions of parameter space, use of a novel type of mean-field approximation leads to a uniformly accurate approximation over all parameter space. In particular, we show that the latter is practically indistinguishable from distributions calculated using the stochastic simulation algorithm. We conclude in Section V by a discussion of our method versus those already in the literature.

II. STOCHASTIC MODEL OF TRANSCRIPTION KINETICS

In this Section we define the model and introduce the master equation that describes its stochastic evolution in time. While as mentioned in the Introduction, the model has many interpretations, in what follows we will describe it in terms of RNAP dynamics during transcription.

A. Definition of the model

Transcription is modelled as a series of steps involving promoter activation, initiation at the transcription start site, RNAP movement along the gene and termination at the stop site that frees the newly synthesized mRNA.

The promoter is modelled having two states, activated (on state G) and deactivated (off state G^*). Gene body—the part of the gene between the start and stop sites—is divided into L segments. The state of the system is described by the state of the promoter (on or off), and the number of RNA polymerases in each of the L segments, n_i for $i = 1, \dots, L$.

The model is summarised in Fig. 1 – note that this is a cartoon version of the reaction scheme (1) which is specific to the context of transcription. In particular the RNAP on gene segment i is what was labelled as P_i in scheme (1). Note that the number of RNAPs on a gene is equal to the number of nascent mRNA; this is because each RNAP has attached to it a nascent mRNA tail whose length increases as elongation proceeds. The rates of gene activation and deactivation are s_u and s_b, respectively. Once the promoter is active (in the on state), transcription initiation occurs at rate r, whereby a new RNAP enters the first segment. We do not consider leaky transcription, i.e. transcription initiation is not allowed when the promoter is in the off state. After initiation, the RNA polymerese moves along the gene body from segment to segment with, in general, segment-dependent elongation rate $k_{\mathrm{el},i}$. During elongation, the RNAP can prematurely detach from the gene at rate d. Transcription termination occurs at the last segment with rate $k_{\mathrm{el},L}$, whereby the RNAP is removed from the gene and the newly synthesized mRNA is released.

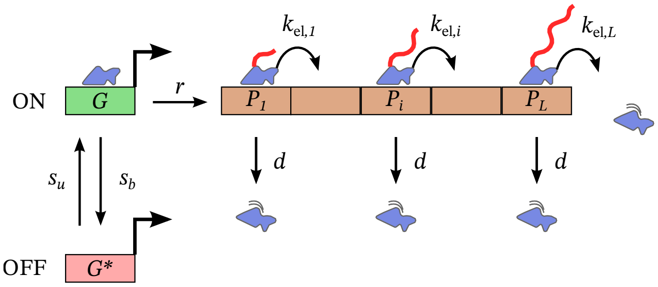

FIG. 1. Schematic of the stochastic transcription model. The gene body is divided into L segments. The promoter can be in two states: active (on) and inactive (off). Promoter on and off switching rates are s_u and s_b, respectively. Transcription initiation occurs when the promoter is in the on state at rate r by which a new RNAP is added to the first segment. No transcription initiation is allowed when the promoter is in the off state. Following initiation, RNAPs move along the gene with segment-dependent elongation rates $k_{\mathrm{el},i}$. They can detach from DNA at any segment with rate d (premature termination). Once they reach the stop site they terminate transcription at rate $k_{\mathrm{el},L}$. The model ignores excluded volume interactions between neighbouring RNAPs, i.e. the number of RNAPs in each segment is unbounded.

We note that the model does not explicitly take into account excluded volume interactions between RNAPs. RNAP has a footprint of $\ell_{RNAP} = 35$ base pairs (bp) [20]. The size of each segment is $\ell_{\mathrm{segment}} = L_{\mathrm{gene}}/L$, where L_{gene} is the number of base pairs in the gene body, typically measured in thousands of base pairs. Thus the maximum number of RNAPs in each segment is $c = \ell_{\mathrm{segment}}/\ell_{\mathrm{RNAP}}$, whereas in our model the number of RNAPs in each segment is unbounded.

As we show later, this simplification allows us to obtain analytical expressions for the distribution of RNAPs along the gene. In order to compensate for this simplification, one can choose the size of the segment ℓ_{segment} such that the average number of RNAPs in any segment is less than the maximum capacity c

$$\langle n_i \rangle < c, \quad i = 1, \dots, L. \tag{2}$$

Alternatively, one can choose ℓ_{segment} such that the probability of observing more than c RNAPs in each segment is small,

$$P(n_i > c) \ll 1, \quad i = 1, \ldots, L. \tag{3}$$

We provide analytical expression for both $\langle n_i \rangle$ and $P(n_i)$ so that any of these two conditions can be easily checked if the model parameters are known. Transcription is typically rate-limited by initiation, which means there are only few active RNAPs on the gene at a given time, and hence conditions (2) and (3) are automatically satisfied.

B. Master equations for the RNAP distribution

The central information that we are interested in is the probability $P(n_1, \ldots, n_L; t)$ to find n_1 RNAPs in segment 1, n_2 RNAPs in segment 2, and so on, at a given time t. This probability is in fact a sum of two probabilities $P_{\text{on}}(n_1, \ldots, n_L; t)$ and $P_{\text{off}}(n_1, \ldots, n_L; t)$ which are conditioned on the state of the promoter. The probabilities P_{on} and P_{off} satisfy the following master equations,

$$\begin{aligned}\frac{\partial}{\partial t} P_{\text{on}}(n_1, \ldots, n_L; t) &= \sum_{i=1}^{L-1} k_{\text{el},i}(n_i+1)\mathbb{E}_i \mathbb{E}_{i+1}^{-1} P_{\text{on}} \\ &+ k_{\text{el},L}(n_L+1)\mathbb{E}_L P_{\text{on}} + \sum_{i=1}^{L} d(n_i+1)\mathbb{E}_i P_{\text{on}} \\ &+ r\mathbb{E}_1^{-1} P_{\text{on}} + s_u P_{\text{off}} - \sum_{i=1}^{L}(k_{\text{el},i}+d) n_i P_{\text{on}} \\ &- (s_b + r) P_{\text{on}},\end{aligned} \tag{4a}$$

$$\begin{aligned}\frac{\partial}{\partial t} P_{\text{off}}(n_1, \ldots, n_L; t) &= \sum_{i=1}^{L-1} k_{\text{el},i}(n_i+1)\mathbb{E}_i \mathbb{E}_{i+1}^{-1} P_{\text{off}} \\ &+ k_{\text{el},L}(n_L+1)\mathbb{E}_L P_{\text{off}} + \sum_{i=1}^{L} d(n_i+1)\mathbb{E}_i P_{\text{off}} \\ &+ s_b P_{\text{on}} - \sum_{i=1}^{L}(k_{\text{el},i}+d) n_i P_{\text{off}} - s_u P_{\text{off}}.\end{aligned} \tag{4b}$$

Here we have shortened the notation by introducing step operators $\mathbb{E}_i$ and $\mathbb{E}_i^{-1}$ [21] that increase and decrease the number of particles n_i in segment i, respectively,

$$\mathbb{E}_i P(\ldots, n_i, \ldots) = P(\ldots, n_i+1, \ldots), \tag{5a}$$

$$\mathbb{E}_i^{-1} P(\ldots, n_i, \ldots) = P(\ldots, n_i-1, \ldots), \quad n_i \geq 1, \tag{5b}$$

$$\mathbb{E}_i^{-1} P(\ldots, n_i = 0, \ldots) = 0. \tag{5c}$$

From now on, we are interested only in the steady state, so that $\partial P_{\text{on}}/\partial t = 0$ and $\partial P_{\text{off}}/\partial t = 0$ and so we drop the time dependence from P_{on} and P_{off}.

Rather than working with the master equation directly, we introduce the generating functions G_{on}, G_{off} for each of the promoter states, and also $G = G_{\text{on}} + G_{\text{off}}$, whereby

$$\begin{aligned}G_{\text{on}}(z_1, \ldots, z_L) &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_L=0}^{\infty} z_1^{n_1} \ldots z_L^{n_L} \\ &\times P_{\text{on}}(n_1, \ldots, n_L),\end{aligned} \tag{6a}$$

$$\begin{aligned}G_{\text{off}}(z_1, \ldots, z_L) &= \sum_{n_1=0}^{\infty} \cdots \sum_{n_L=0}^{\infty} z_1^{n_1} \ldots z_L^{n_L} \\ &\times P_{\text{off}}(n_1, \ldots, n_L).\end{aligned} \tag{6b}$$

From the master equation we get the following steady-state partial differential equations for G_{on} and G_{off},

$$\begin{aligned}&\sum_{i=1}^{L-1} \Big[(k_{\text{el},i}+d)z_i - k_{\text{el},i} z_{i+1} - d\Big] \frac{\partial G_{\text{on}}}{\partial z_i} \\ &- (k_{\text{el},L}+d)(1-z_L)\frac{\partial G_{\text{on}}}{\partial z_L} = s_u G_{\text{off}} - s_b G_{\text{on}} \\ &- r(1-z_1)G_{\text{on}},\end{aligned} \tag{7a}$$

$$\begin{aligned}&\sum_{i=1}^{L-1} \Big[(k_{\text{el},i}+d)z_i - k_{\text{el},i} z_{i+1} - d\Big] \frac{\partial G_{\text{off}}}{\partial z_i} \\ &- (k_{\text{el},L}+d)(1-z_L)\frac{\partial G_{\text{off}}}{\partial z_L} = s_b G_{\text{on}} - s_u G_{\text{off}}.\end{aligned} \tag{7b}$$

To the best of our knowledge, this system of equations has not been solved in closed-form before, except in three special cases: telegraph model ($L = 1$), constitutive gene expression ($s_b = 0$) and deterministic elongation with mature RNA production. For completeness we summarise these three cases below.

C. Exactly solvable cases

1. The telegraph model

The case of $L = 1$ is known as the telegraph model [5], whose solution for the generating function reads

$$G(z_1) = e^{-\alpha(1-z_1)} M(\sigma_b; \sigma_u + \sigma_b; \alpha(1-z_1)), \tag{8}$$

where $k_{\text{el},1} = k_{\text{el}}$ and

$$\alpha = \frac{r}{k_{\text{el}}+d}, \quad \sigma_u = \frac{s_u}{k_{\text{el}}+d}, \quad \sigma_b = \frac{s_b}{k_{\text{el}}+d}. \tag{9}$$

In the expression for $G(z_1)$, M is Kummer's (confluent hypergeometric) function [22]. The probability distribution of n_1 is given by

$$\begin{aligned}P(n_1) &= \frac{\alpha^{n_1} e^{-\alpha}}{n_1!} \frac{(\sigma_u)_{n_1}}{(\sigma_u+\sigma_b)_{n_1}} \\ &\times M(\sigma_b, \sigma_u + \sigma_b + n_1, \alpha).\end{aligned} \tag{10}$$

We note this expression also holds for the marginal distribution $P(n_1)$ of the model with $L \geq 1$ – this is because the fluctuations in the RNAP numbers on segment 1 are unaffected by the fluctuations on the other segments due to the irreversible nature of the reactions between segments.

2. *Constitutive gene expression*

A constitutive promoter is always in the on state, which means that $s_b = 0$. In that case there is only one equation to solve, that for $G_{\mathrm{on}} \equiv G$, since $G_{\mathrm{off}} = 0$. We can easily check that the solution is given by

$$G(z_1, \dots, z_L) = \prod_{i=1}^{L} e^{-\mu_i(1-z_i)}, \tag{11}$$

where μ_i is the average number of RNAPs in segment i and is given by Eq. (22) in which $\eta = s_u/(s_u + s_b) = 1$. This generating function leads to a probability distribution that is a product of Poisson distributions,

$$P(n_1, \dots, n_L) = \prod_{j=1}^{L} \frac{\mu_j^{n_j} e^{-\mu_j}}{n_j!}. \tag{12}$$

Consequently there are no correlations between segments. Models of this type have been used to describe mRNA senescence [23] and splicing [24].

3. *Deterministic elongation with mature RNA production*

We can extend this model to include mature RNA production by another stage after transcription termination. The model now has $L+1$ segments, whereby the first L segments describe RNAP dynamics and the last segment counts the number of mature RNA m. We further assume uniform elongation rates in segments $i = 1, \dots, L$ so that $k_{el,i} = k_{\mathrm{el}}$. At the last segment, the mature RNA is degraded at rate d_m, which in our notation is equivalent of saying that $d_m \equiv k_{\mathrm{el},L+1} + d$.

For this model the distribution of mature RNA, $P(m)$, has been found analytically [16] in the limit in which the elongation is deterministic, i.e when $d_m/k_{\mathrm{el}} \to 0$ and $L \to \infty$ but keeping the transcription time $T = L/k_{\mathrm{el}}$ fixed. In this limit, the distribution of the mature RNA is the same as in the telegraph model; hence this limit is a more formal way of deriving the telegraph model from a detailed model of RNAP dynamics.

III. EXACT MOMENTS OF THE RNAP DISTRIBUTION P

In this Section we consider the first three moments of the RNAP distribution P in the steady state. We will need these moments later in Section IV, where we derive an analytical expression for the RNAP distribution using mean-field theory.

We begin by deriving recurrence relations for the first three moments, which we solve analytically in the uniform case (all $k_{\mathrm{el},i}$ are equal to k_{el}). In the non-uniform case (arbitrary values of $k_{\mathrm{el},i}$ in each segment) we obtain the first moments analytically and the second and third moments numerically.

In the former case, the first two moments have been previously derived in Ref. [16]. Here for the uniform case we present a new derivation using lattice paths which has the advantage that it can be extended to higher moments, as we demonstrate by computing the third moment.

We first make the following change of variables,

$$u_i = \lambda_i(1 - z_i), \tag{13}$$

where λ_i is given by

$$\lambda_1 = 1, \quad \lambda_i = \prod_{j=1}^{i-1} \frac{k_{\mathrm{el},j}}{k_{\mathrm{el},j} + d}, \quad i = 2, \dots, L. \tag{14}$$

The equations for G_{on} and G_{off} now take a simpler form,

$$\sum_{i=1}^{L-1} \omega_i(u_i - u_{i+1}) \frac{\partial G_{\mathrm{on}}}{\partial u_i} + \omega_L u_L \frac{\partial G_{\mathrm{on}}}{\partial u_L} = s_u G_{\mathrm{off}} - (s_b + r u_1)\, G_{\mathrm{on}}, \tag{15a}$$

$$\sum_{i=1}^{L-1} \omega_i(u_i - u_{i+1}) \frac{\partial G_{\mathrm{off}}}{\partial u_i} + \omega_L u_L \frac{\partial G_{\mathrm{off}}}{\partial u_L} = s_b G_{\mathrm{on}} - s_u G_{\mathrm{off}}, \tag{15b}$$

where we have introduced ω_i defined as

$$\omega_i = k_{\mathrm{el},i} + d. \tag{16}$$

The recurrence relations for the moments of P can be found by taking partial derivatives of the equations above and setting $u_1 = \dots = u_L = 0$.

A. First moments of P

We introduce the following notation,

$$g^{(i)} = g^{(i)}_{\mathrm{on}} + g^{(i)}_{\mathrm{off}}, \tag{17}$$

where

$$g^{(i)}_{\mathrm{on}} = \left.\frac{\partial G_{\mathrm{on}}}{\partial u_i}\right|_{\{u_j\}=0}, \quad g^{(i)}_{\mathrm{off}} = \left.\frac{\partial G_{\mathrm{off}}}{\partial u_i}\right|_{\{u_j\}=0}, \tag{18}$$

and $\{u_j\} = 0$ means that $u_1 = \dots = u_L = 0$. In order to find $g^{(i)}$ we add Eqs. (15a) and (15b) together, then take a partial derivative with respect to u_i and set all $u_1 = \dots = u_L = 0$. The resulting recurrence relation for $g^{(i)}$ reads

$$\omega_i g^{(i)} = \omega_{i-1} g^{(i-1)}, \quad \omega_1 g^{(1)} = -r\eta, \tag{19}$$

where η is the fraction of the time that the promoter spends in the on state,

$$\eta = \frac{s_u}{s_u + s_b}. \tag{20}$$

From here we get

$$g^{(i)} = -\frac{r\eta}{\omega_i}. \tag{21}$$

The first moment of P in segment i is thus given by

$$\mu_i = \langle n_i \rangle = -\lambda_i g^{(i)} = \frac{r\eta\lambda_i}{\omega_i} = \frac{r\eta}{k_{\mathrm{el},i}} \prod_{j=1}^{i} \frac{k_{\mathrm{el},j}}{k_{\mathrm{el},j} + d}. \tag{22}$$

For uniform rates with no detachment, the mean is same across the gene. Otherwise the mean varies across the gene, for example if the elongation rates are non-uniform with no detachment or if the elongation rates are uniform with non-zero detachment

We also write down an analytical expression for $g_{\mathrm{on}}^{(i)}$ because we will need it later for computing the second moments of P,

$$g_{\mathrm{on}}^{(i)} = -\frac{r\eta}{\omega_i} \sum_{n=1}^{i} \left(\delta_{n,1} + \frac{s_u}{\omega_n} \right) \times \prod_{k=n}^{i} \frac{\omega_k}{s_u + s_b + \omega_k}, \tag{23}$$

where $\delta_{i,j}$ is the Kronecker delta function. In the case of uniform elongation rates, the expression for $g_{\mathrm{on}}^{(i)}$ takes a simpler form:

$$g_{\mathrm{on}}^{(i)} = -\alpha\eta \left(\eta + \frac{1-\eta}{(1+\sigma_u+\sigma_b)^i} \right), \tag{24}$$

where the rescaled parameters α, σ_u and σ_b are defined in Eq. (9).

B. Second moments of P

In order to find the second moments of P we define

$$g^{(ij)} = g_{\mathrm{on}}^{(ij)} + g_{\mathrm{off}}^{(ij)}, \tag{25}$$

where

$$g_{\mathrm{on}}^{(ij)} = \left. \frac{\partial G_{\mathrm{on}}^2}{\partial u_i \partial u_j} \right|_{\{u_j\}=0}, \quad g_{\mathrm{off}}^{(ij)} = \left. \frac{\partial G_{\mathrm{off}}^2}{\partial u_i \partial j} \right|_{\{u_j\}=0}. \tag{26}$$

Again, we add Eqs. (15a) and (15b) together, then take partial derivatives with respect to u_i and u_j and set $u_1 = \cdots = u_L = 0$, yielding the following recurrence relations for $g^{(ij)}$,

$$g^{(ij)} = \frac{\omega_{i-1}}{\omega_i + \omega_j} g^{(i-1j)} + \frac{\omega_{j-1}}{\omega_i + \omega_j} g^{(ij-1)}, \quad i, j = 2, \ldots, L \tag{27a}$$

$$g^{(1j)} = \frac{\omega_{j-1}}{\omega_1 + \omega_j} g^{(1j-1)} - \frac{r g_{\mathrm{on}}^{(j)}}{\omega_1 + \omega_j}, \; j = 2, \ldots, L \tag{27b}$$

$$g^{(i1)} = \frac{\omega_{i-1}}{\omega_i + \omega_1} g^{(i-11)} - \frac{r g_{\mathrm{on}}^{(i)}}{\omega_i + \omega_1}, \; i = 2, \ldots, L \tag{27c}$$

$$g^{(11)} = -\frac{r}{\omega_1} g_{\mathrm{on}}^{(1)}. \tag{27d}$$

A closed-form expression for $g^{(ij)}$ can be found in the case of uniform elongation rates using lattice path methods as we show below. In the non-uniform case it is possible to solve the recurrence relations directly (as lattice path methods cannot be used), but it is simpler to solve them numerically. Once we find $g^{(ij)}$, the covariance $\mathrm{cov}(n_i, n_j)$ can be found from

$$\begin{aligned} \mathrm{cov}(n_i, n_j) &= \langle (n_i - \mu_i)(n_j - \mu_j) \rangle \\ &= \lambda_i \lambda_j (g^{(ij)} - g^{(i)} g^{(j)}) - \delta_{i,j} \lambda_i g^{(i)}. \end{aligned} \tag{28}$$

In the case of uniform elongation rates ω_i are all equal and the coefficients in Eqs. (27) become all equal to $1/2$. We can think of Eqs. (27) in terms of two-dimensional lattice paths with unit steps $(m, n) \to (m-1, n)$ and $(m, n) \to (m, n-1)$. We are interested in the paths that start at (i, j) and end at one of the boundary points, $(1, n)$ or $(m, 1)$, for $1 \leq m \leq i$ and $1 \leq n \leq j$. Let us say we start at (i, j) and end at $(m, 1)$ for $1 \leq m \leq i$. Each step we make is weighted by factor $1/2$, until we reach the end point at $(m, 1)$, which is weighted by $(-\alpha g_{\mathrm{on}}^{(m)}/2)$, where $\alpha = r/(k_{\mathrm{el}} + d)$. Similarly, the end point at $(1, n)$ for $1 \leq n \leq j$ is weighted by $(-\alpha g_{\mathrm{on}}^{(n)})/2$, whereas the end point at $(1, 1)$ is weighted by $-\alpha g_{\mathrm{on}}^{(1)}$.

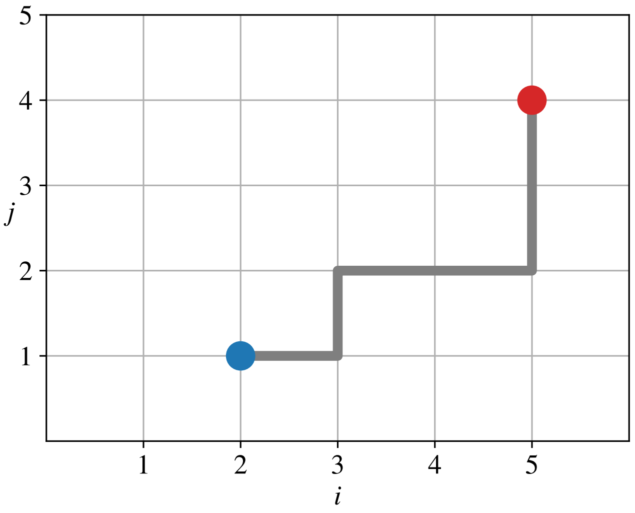

FIG. 2. A lattice path from point $(i, j) = (5, 4)$ (red) to point $(2, 1)$ (blue). The contribution from this path to $g^{(54)}$ is $(1/2)^6(-\alpha g_{\mathrm{on}}^{(2)}/2)$.

One such lattice path is presented in Fig. 2 with the start point at $(5, 4)$ and the end point at $(2, 1)$. The total weight of this path is $1/2^6$—the length of the path is 6—multiplied by the weight at the end point, $(-\alpha g_{\mathrm{on}}^{(2)}/2)$, which gives

$$\text{path weight in Fig. 2} = \left(\frac{1}{2}\right)^6 \left(-\frac{\alpha g_{\rm on}^{(2)}}{2}\right). \tag{29}$$

Because all paths between two given points on the lattice carry the same weight, we can simply count them, multiply their weight by the weight at the boundary and then sum over all boundary points. This simplification is not possible in the case of non-uniform elongation rates because paths between two given points on the lattice will in general all have different weights.

In order to count the paths between two fixed points, we denote by S moving south and by W moving west. We can then describe a path of length n as an n-letter word consisting of letters S and W. For example, the path in Fig. 2 can be written as SSWWSWW.

Next, we count all paths from a given point (i, j) to the boundary point $(1, n)$ for each $n = 2, \ldots, j$. To get to this point we need to make $i-1$ steps to the south and $j-n$ steps to the west. The total length of this path is $i-1+j-n$ and there are $\binom{i-1+j-n}{i-1}$ such paths. We repeat this procedure for all paths from (i, j) to the boundary points $(m, 1)$ for each $m = 2, \ldots, i$. Finally, we count all paths from (i, j) to $(1, 1)$. After rearranging terms, the total contribution from all paths can be written as

$$g^{(ij)} = \sum_{m=1}^{i} \frac{1}{2^{i-m+j-1}} \binom{i-m+j-1}{j-1} \left(-\frac{\alpha g_{\rm on}^{(m)}}{2}\right) + \sum_{n=1}^{j} \frac{1}{2^{i-1+j-n}} \binom{i-1+j-n}{i-1} \left(-\frac{\alpha g_{\rm on}^{(n)}}{2}\right). \tag{30}$$

Inserting the expression for $g_{\rm on}^{(m)}$ from Eq. (24) into (30) we get a closed-form expression for the second moment of P (the covariance)

$$\mathrm{cov}(n_i, n_j) = \delta_{i,j}\mu_i + \mu_i\mu_j \frac{\sigma_b/\sigma_u}{1+\sigma_u+\sigma_b}\gamma_{ij}(\sigma_u, \sigma_b), \tag{31}$$

where $\gamma_{i,j}(\sigma_u, \sigma_b)$[25] is given by

$$\gamma_{ij}(\sigma_u, \sigma_b) = \sum_{q=0}^{i-1} \binom{j-1+q}{q} \frac{1}{2^{j+q}(1+\sigma_u+\sigma_b)^{i-q-1}} + \sum_{q=0}^{j-1} \binom{i-1+q}{q} \frac{1}{2^{i+q}(1+\sigma_u+\sigma_b)^{j-q-1}}. \tag{32}$$

The derivation of the second moments conditioned on the promoter being in the on state can be done is a similar fashion. We omit here the details and state only the final result:

$$g_{\rm on}^{(ij)} = \frac{1}{2+\sigma_u+\sigma_b} \sum_{m=1}^{i}\sum_{n=1}^{j} \binom{i-m+j-n}{i-m} \times \frac{(\sigma_u g^{(mn)} - \delta_{m1}\alpha g_{\rm on}^{(n)} - \delta_{1n}\alpha g_{\rm on}^{(m)})}{(2+\sigma_u+\sigma_b)^{i-m+j-n}}. \tag{33}$$

It is interesting that while RNAPs in this model are non-interacting (dynamics of a single RNAP does not depend on the presence of other RNAPs), yet the model exhibits long-range correlations in RNAP numbers between segments. These correlations are entirely due to promoter switching, because the model with constitutive gene expression ($s_b = 0$) has covariance $\mathrm{cov}(n_i, n_j) = 0$ for $i \neq j$, see Section II C.

C. Third moments of P

In order to compute third moments pf P, we consider third-order partial derivatives of $G_{\rm on}$ and of $G_{\rm off}$,

$$g^{(ijk)} = g_{\rm on}^{(ijk)} + g_{\rm off}^{(ijk)}, \tag{34}$$

where

$$g_{\rm on}^{(ijk)} = \left.\frac{\partial^3 G_{\rm on}}{\partial u_i u_j u_k}\right|_{\{u_j\}=0}, \tag{35a}$$

$$g_{\rm off}^{(ijk)} = \left.\frac{\partial^3 G_{\rm off}}{\partial u_i u_j u_k}\right|_{\{u_j\}=0}. \tag{35b}$$

The recipe to obtain the recurrence relation for $g^{(ijk)}$ is the same as before. We first add Eqs. (15a) and (15b) together, then take the partial derivatives with respect to u_i, u_j and u_k and finally set $u_1 = \cdots = u_L = 0$. The resulting recurrence equations are given by

$$g^{(ijk)} = \frac{\omega_{i-1}}{\omega_i+\omega_j+\omega_k} g^{i-1jk} + \frac{\omega_{j-1}}{\omega_i+\omega_j+\omega_k} g^{ij-1k} + \frac{\omega_{k-1}}{\omega_i+\omega_j+\omega_k} g^{ijk-1}, \quad i,j,k = 2,\ldots,L, \tag{36a}$$

$$g^{(1jk)} = \frac{\omega_{j-1}}{\omega_1+\omega_j+\omega_k} g^{1j-1k} + \frac{\omega_{k-1}}{\omega_1+\omega_j+\omega_k} g^{1jk-1} - \frac{r g_{\rm on}^{(jk)}}{\omega_1+\omega_j+\omega_k} \quad j,k = 2,\ldots,L, \tag{36b}$$

$$g^{(11k)} = \frac{\omega_{k-1}}{2\omega_1+\omega_k} - \frac{2r g_{\rm on}^{(1k)}}{2\omega_1+\omega_k}, \quad k = 2,\ldots,L \tag{36c}$$

$$g^{(111)} = -r g_{\rm on}^{(11)}. \tag{36d}$$

Other equations, for example for $g^{(i1k)}$, can be obtained from these equations using that fact $g^{(ijk)}$ is invariant under the permutation of indices i, j, k. Eqs. (36a) can be solved analytically in the case of uniform elongation rates using lattice paths, or numerically in the non-uniform case. In the former case we consider three-dimensional lattice paths that start at (i, j, k) and end at one of the boundary points $(1, n, o)$, $(m, 1, o)$ or $(m, n, 1)$, where $1 \leq m \leq i$, $1 \leq n \leq j$ and $1 \leq o \leq k$. Instead of $1/2$, each step is now weighted by $1/3$. If the end point is, say $(1, n, o)$, where $n, o \neq 1$, then the weight of that point is $-\alpha g_{\rm on}^{(no)}/3$, where $g_{\rm on}^{(no)}$ is the second moment conditioned on the promoter being in the on state. If the end point is

$(1,1,o)$ for $o \neq 1$, then the weight is $-2\alpha g_{\mathrm{on}}^{(1o)}/3$. If $o=1$ then the weight is $-\alpha g_{\mathrm{on}}^{(no)}$. The final result for $g^{(ijk)}$ is

$$\begin{aligned} g^{(ijk)} = &\sum_{m=1}^{i}\sum_{n=1}^{j}\binom{i-m+j-n+k-1}{i-m,\, j-n,\, k-1} \\ &\times \frac{1}{3^{i-m+j-n+k-1}}\left(-\frac{\alpha g_{\mathrm{on}}^{(mn)}}{3}\right) \\ &+\sum_{m=1}^{i}\sum_{o=1}^{k}\binom{i-m+j-1+k-o}{i-m,\, j-1,\, k-o} \\ &\times \frac{1}{3^{i-m+j-1+k-o}}\left(-\frac{\alpha g_{\mathrm{on}}^{(mo)}}{3}\right) \\ &+\sum_{n=1}^{j}\sum_{o=1}^{k}\binom{i-1+j-n+k-o}{i-1,\, j-n,\, k-o} \\ &\times \frac{1}{3^{i-1+j-n+k-o}}\left(-\frac{\alpha g_{\mathrm{on}}^{(no)}}{3}\right) \end{aligned} \tag{37}$$

where $g_{\mathrm{on}}^{(ij)}$ is given by Eq. (33).

Of particular importance is the third standardised central moment in segment i, which is given by

$$\begin{aligned} \kappa_i &= \frac{\langle (n_i - \langle n_i \rangle)^3 \rangle}{\sigma_i^3} \\ &= \frac{1}{\sigma_i^3}\Big(-\lambda_i^3 (g^{(iii)} - 3g^{(ii)}g^{(i)} + 2g^{(i)}) \\ &\quad + 3\lambda_i^2\left[g^{(ii)} - \left(g^{(i)}\right)^2\right] - \lambda_i g^{(i)}\Big), \end{aligned} \tag{38}$$

where $g^{(i)}$, $g^{(ij)}$ and $g^{(ijk)}$ are given by Eqs. (21), (30) and (37), respectively, and σ_i is the standard deviation of n_i,

$$\sigma_i = \sqrt{\langle (n_i - \langle n_i \rangle)^2 \rangle} = \sqrt{\mathrm{cov}(n_i, n_i)}. \tag{39}$$

The third standardised central moment or skewness is useful for understanding the shape of the distribution, in particular its (a)symmetry. We will use this expression later in Section IV.

IV. THE RNAP DISTRIBUTION IN THE MEAN-FIELD APPROXIMATION

As we mentioned earlier, solving the system of partial differential equations for G_{on} and G_{off} for arbitrary number of segments L is a difficult problem. In order to make progress, we focus on the marginal distribution of n_i in the steady state defined as

$$P(n_i) = \sum_{\{n_i\}\setminus n_i} P(n_1, \ldots, n_L), \tag{40}$$

where the summation goes over all values of variables $n_1, \ldots, n_L$ except for the variable n_i which is kept fixed. In this section we provide an analytical expression for $P(n_i)$ in the mean-field approximation.

A. Naive mean-field approximation

We first consider an approximation that we call the naive mean-field approximation (NMF). This type of approximation is well known in statistical physics and often gives satisfactory results in the absence of long-range correlations (for example, away from phase transitions). However, we will show that for this model the naive mean-field approximation does not always work, and later we find an improved mean-field approximation that shows excellent agreement with the exact results from stochastic simulations.

The main idea of the naive mean-field approximation is to find an effective master equation for the marginal distribution $P(n_i)$ that is decoupled from fluctuations in segment $i-1$ in which case we can solve it analytically. We do this by summing Eqs. (4a) and (4b) over all n_j for all $j \neq i$. Many terms cancel in the summation and we are left with master equations for $P_{\mathrm{on}}(n_i)$ and $P_{\mathrm{off}}(n_i)$. In each master equation there will be only one term that couples segment i with $i-1$, and that is the term that describes elongation from $i-1$ to i. For example, in the master equation for P_{on}, this term is given by

$$\begin{aligned} &\sum_{n_{i-1}=0}^{\infty} k_{\mathrm{el},i-1}(n_{i-1}+1)P_{\mathrm{on}}(n_{i-1}+1, n_i-1) \\ &= \left(k_{\mathrm{el},i-1}\sum_{n_{i-1}=0}^{\infty} n_{i-1}P_{\mathrm{on}}(n_{i-1}|n_i-1)\right) \\ &\quad \times P_{\mathrm{on}}(n_i-1). \end{aligned} \tag{41}$$

The expression in the parentheses is the effective rate at which new RNAPs are added to segment i from segment $i-1$. The problem is of course that we cannot compute this rate unless we compute the conditional distribution $P_{\mathrm{on}}(n_{i-1}|n_i-1)$. The naive mean-field approximation amounts to replacing

$$P_{\mathrm{on}}(n_{i-1}|n_i-1) \approx \frac{P_{\mathrm{on}}(n_{i-1})}{\sum_{n_{i-1}} P_{\mathrm{on}}(n_{i-1})}, \tag{42a}$$

$$P_{\mathrm{off}}(n_{i-1}|n_i-1) \approx \frac{P_{\mathrm{off}}(n_{i-1})}{\sum_{n_{i-1}} P_{\mathrm{off}}(n_{i-1})}. \tag{42b}$$

In other words we ignore correlations between segments $i-1$ and i, except for the fact that both segments have the same promoter state. The effective rates p_i and q_i at which new RNAPs are added to segment i in the on and off state are thus given by

$$\begin{aligned} p_i &= \frac{k_{\mathrm{el},i-1}\sum_{n_{i-1}} n_{i-1}P_{\mathrm{on}}(n_{i-1})}{\sum_{n_{i-1}} P_{\mathrm{on}}(n_{i-1})} \\ &= \frac{-k_{\mathrm{el},i-1}\,\lambda_{i-1}g_{\mathrm{on}}^{(i-1)}}{\eta}, \quad i=2,\ldots,L, \end{aligned} \tag{43a}$$

$$q_i = \frac{k_{\mathrm{el},i-1}\sum_{n_{i-1}} n_{i-1}P_{\mathrm{off}}(n_{i-1})}{\sum_{n_{i-1}} P_{\mathrm{off}}(n_{i-1})}$$

$$= \frac{-k_{\mathrm{el},i-1}\, g_{\mathrm{off}}^{(i-1)}}{1-\eta}, \quad i = 2, \ldots, L. \tag{43b}$$

The effective model that governs the stochastic dynamics of n_i in the naive mean-field approximation is equivalent to the telegraph model for RNA production [5] in which transcription rates in the on and off states are p_i and q_i, respectively, and the mRNA degradation rate is $k_{\mathrm{el},i} + d$. The reactions for this model with their corresponding rates are given by

$$G \underset{s_u}{\overset{s_b}{\rightleftharpoons}} G^*, \tag{44a}$$

$$G \xrightarrow{p_i} G + P_i, \tag{44b}$$

$$G^* \xrightarrow{q_i} G^* + P_i, \tag{44c}$$

$$P_i \xrightarrow{k_{\mathrm{el},i}+d} \emptyset, \tag{44d}$$

where G and G^* represent the on and off states of the gene and P_i is the species associated with the i^{th} segment (see scheme (1)). We note that the telegraph model in which transcription is allowed from both on and off states is also known as the leaky telegraph model [11].

The telegraph model described by reactions (44a)-(44d) can be solved exactly and the generating function reads

$$G_0(z_i) = \sum_{n_i=0}^{\infty} z_i^{n_i} P_0(n_i) = e^{-\alpha_i(1-z_i)} \times M(\sigma_b; \sigma_u + \sigma_b; (\alpha_i - \beta_i)(1-z_i)), \tag{45}$$

where α_i, β_i, $\sigma_{u,i}$ and $\sigma_{b,i}$ are defined as

$$\alpha_i = \frac{p_i}{k_{el,i}+d}, \quad \beta_i = \frac{q_i}{k_{el,i}+d}, \quad \sigma_{u,i} = \frac{s_u}{k_{el,i}+d}, \quad \sigma_{b,i} = \frac{s_b}{k_{el,i}+d}. \tag{46}$$

The probability distribution $P(n_i)$ in the naive mean-field approximation reads

$$P(n_i) = \frac{e^{-\alpha_i}}{n_i!} \sum_{k=0}^{n_i} \binom{n_i}{k} \beta_i^{n_i-k}(\alpha_i-\beta_i)^k \times \frac{(\sigma_{u,i})_k}{(\sigma_{u,i}+\sigma_{b,i})_k} M(\sigma_{b,i}; \sigma_{u,i}+\sigma_{b,i}+k; (\alpha_i-\beta_i)), \tag{47}$$

and the mean and the variance are given by, respectively,

$$\mu_{0,i} = \alpha_i\eta + \beta_i(1-\eta), \tag{48}$$

$$\sigma_{0,i}^2 = \mu_{0,i} + (\alpha_i-\beta_i)^2 \frac{\eta(1-\eta)}{1+\sigma_{u,i}+\sigma_{b,i}}. \tag{49}$$

We can check that the choice of p_i and q_i in Eqs. (43a) and (43b) ensures that $\mu_{0,i} = \mu_i$, i.e. the mean of n_i computed in Eq. (22) is the same as the mean in the naive mean-field approximation. However the value of the variance is generally different from the exact value.

In the case of uniform elongation rate the expression for the variance simplifies to

$$\sigma_{0,i}^2 = \mu_i + \mu_i^2 \frac{\sigma_b/\sigma_u}{1+\sigma_u+\sigma_b} \gamma_{0,ii}(\sigma_u, \sigma_b), \tag{50}$$

where $\gamma_{0,ii}(\sigma_u, \sigma_b)$ is given by

$$\gamma_{0,ii}(\sigma_u, \sigma_b) = \frac{1}{(1+\sigma_u+\sigma_b)^{2i-2}}. \tag{51}$$

The difference between the exact variance in Eq. (31) and Eq. (50) is in the factor γ_{ii} for $i \geq 2$ (the naive mean-field approximation is exact in segment 1). In general we find that $\gamma_{0,ii} < \gamma_{ii}$ for any $i = 2, \ldots, L$ for $\sigma_u \neq 0$ and $\sigma_b \neq 0$, whereas $\gamma_{0,ii} = \gamma_{ii} = 1$ for $\sigma_u = \sigma_b = 0$. The naive mean-field approximation underestimates the exact variance, which is expected given that we ignored correlations. When $\sigma_b = 0$ (constitutive gene expression), there are no correlations, the naive mean-field approximation becomes exact and the resulting distribution is Poisson, see Eq. (12).

In general, the naive mean-field approximation is valid if the correlations between n_{i-1} and n_i are small. Fortunately, for this model we can compute these correlations exactly from the covariance $\mathrm{cov}(n_{i-1}, n_i)$ in Eq. (28). In particular we can compute the correlation coefficient defined as

$$R_i = \frac{\mathrm{cov}(n_{i-1}, n_i)}{\sigma_{i-1}\sigma_i}, \tag{52}$$

where σ_i is the standard deviation of n_i, see Eq. (39). The naive mean-field is valid provided $R_i \approx 0$, i.e. when n_{i-1} and n_i are not correlated. Computing R_i thus gives us a direct method to check the validity of the naive mean-field theory.

In Fig. 3 we compare RNAP number distribution in the naive mean-field approximation with the exact distribution obtained from stochastic simulations, for two segments on the gene, $i = 2$ and $i = L = 30$. In the case of constitutive gene expression (Fig. 3(a)) we find an excellent agreement between the distributions, in accordance with the fact $R_2 = 1.5 \cdot 10^{-3}$ and $R_{30} = 10^{-4}$. In the case of bursty expression the agreement is not so good for $i = 2$, but improves for $i = L$ (Fig. 3(b)). Again, this can be explained by stronger correlations at the start of the gene than at the end, $R_2 = 0.35$ and $R_{30} = 0.18$. The disagreement between the distributions is strongest in the bimodal and nearly bimodal cases (Figs. 3(c) and (d)). In those cases the correlations are the strongest: we find $R_2 = 0.74$ and $R_{30} = 0.58$ for the bimodal and $R_2 = 0.76$ and $R_{30} = 0.59$ for the nearly bimodal distributions.

B. Improved mean-field approximation

We saw in the previous section that the naive mean-field approximation is equivalent to a telegraph model that produces exact mean but wrong variance. In this

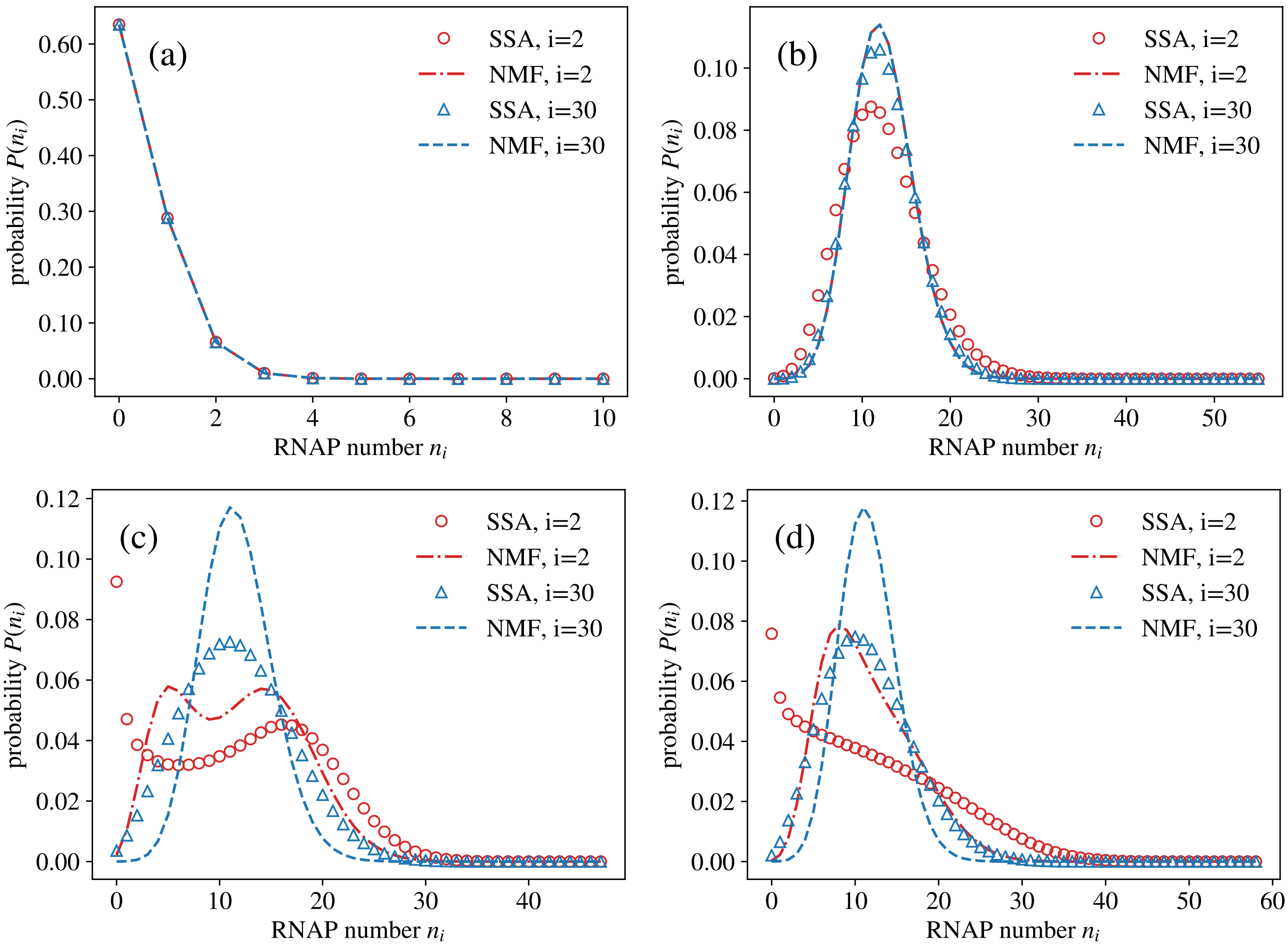

FIG. 3. RNAP number distribution in segments $i = 2$ and $i = L = 30$ in the naive mean-field approximation (NMF; the distribution is given by Eq. (47)), compared to the exact distribution obtained by using the stochastic simulation algorithm (SSA). Elongation rates are assumed to be uniform ($k_{\mathrm{el},i} = k_{\mathrm{el}}$). (a) Constitutive expression: $r = 0.5$, $k_{\mathrm{el}} = 1$, $d = 0$, $s_u = 10$, $s_b = 1$. (b) Bursty expression: $r = 100$, $k_{\mathrm{el}} = 1$, $d = 0$, $s_u = 7$, $s_b = 50$. (c) Bimodal distribution: $r = 20$, $k_{\mathrm{el}} = 1$, $d = 0$, $s_u = 0.35$, $s_b = 0.25$. (d) Nearly bimodal distribution: $r = 30$, $k_{\mathrm{el}} = 1$, $d = 0$, $s_u = 0.5$, $s_b = 0.8$.

section we consider a telegraph model with general rates,

$$G \underset{s'_{u,i}}{\overset{s'_{b,i}}{\rightleftharpoons}} G^*, \tag{53a}$$

$$G \xrightarrow{p'_i} G + P_i, \tag{53b}$$

$$G^* \xrightarrow{q'_i} G^* + P_i, \tag{53c}$$

$$P_i \xrightarrow{k'_{\mathrm{el},i}+d'_i} \emptyset. \tag{53d}$$

We want to find rates p'_i, q'_i, $s'_{u,i}$, $s'_{b,i}$, $k'_{\mathrm{el},i}$ and d'_i such that the mean μ'_i and variance σ'_i of the telegraph model described by Eqs. (53a)-(53d) are both matched to their exact values μ_i in Eq. (22) and $\sigma_i^2 = \mathrm{cov}(n_i, n_i)$ in Eq. (28), respectively. In particular, we want to solve

$$\mu'_i = \alpha'_i \eta' + \beta'_i(1-\eta') = \mu_i, \tag{54a}$$

$$\sigma'_i = \mu'_i + (\alpha'_i - \beta'_i)^2 \frac{\eta'(1-\eta')}{1+\sigma'_{u,i}+\sigma'_{b,i}} = \sigma_i, \tag{54b}$$

where α'_i, β'_i, $\sigma'_{u,i}$ and $\sigma'_{b,i}$ are obtained as before by rescaling p'_i, q'_i, $s'_{u,i}$, $s'_{b,i}$ by $k'_{\mathrm{el},i}+d'_i$, and $\eta'_i = \sigma'_{u,i}/(\sigma'_{u,i}+\sigma'_{b,i})$.

Obviously we cannot find unique rates that solve Eqs. (54a)-(54b) because the problem is overdetermined. We therefore keep the elongation rate $k'_{el,i}$ and the detachment rate d'_i the same as in the original model,

$$k'_{\mathrm{el},i} = k_{\mathrm{el},i}, \quad d'_i = d. \tag{55}$$

That leaves us with four parameters to adjust by solving two equations, Eqs. (54a) and (54b). We further require that the fraction of time the gene spends in the on state, η'_i, is preserved,

$$\eta'_i = \eta. \tag{56}$$

We consider three options for the remaining parameters:

- Option 1: We set $s'_{u,i} = s_u$ and $s'_{b,i} = s_b$ so that Eq. (56) is automatically satisfied. We then adjust the effective transcription rates p'_i and q'_i to match the mean and variance. This is a leaky telegraph model with the same gene switching rates as in the original model.
- Option 2: We set $q'_i = 0$ and adjust p'_i, $s'_{u,i}$ and $s'_{b,i}$ by solving Eqs. (54a), (54b) and (56). This is a non-leaky telegraph model with effective, segment-dependent gene switching rates.
- Option 3: We adjust all four rates p'_i, q'_i, $s'_{u,i}$, $s'_{b,i}$ by solving Eqs. (54a), (54b) and (56). Additionally we require that the skewness predicted by the telegraph model, Eq. (59), is matched to the exact skewness computed in Eq. (38). This is a leaky telegraph model with effective, segment-dependent gene switching rates.

Surprisingly, option 1 does not lead to a noticeable improvement over the naive mean-field approximation, and we discard it. In contrast, option 2 significantly improves the naive mean-field approximation in all four cases: the constitutive, bursty, bimodal and nearly bimodal, as we demonstrate below. Interestingly, option 3 does not provide noticable improvement over option 2. We will show later that that is because option 2 predicts skewness that is very close to the exact one. Since option 2 has a simpler expression for the probability density, we from now on consider only option 2 to which we refer to as the improved mean-field approximation (IMF).

For option 2 the solution to Eqs. (54a), (54b) and (56) with $\beta'_i = 0$ is unique and reads,

$$\alpha'_i = \frac{p'_i}{k_{\mathrm{el},i} + d} = \frac{\mu_i}{\eta}, \tag{57a}$$

$$\sigma'_{u,i} = \frac{s'_{u,i}}{k_{\mathrm{el},i} + d} = \eta \left(\frac{\mu_i^2}{\sigma_i^2 - \mu_i} \frac{1-\eta}{\eta} - 1 \right), \tag{57b}$$

$$\sigma'_{b,i} = \frac{s'_{b,i}}{k_{\mathrm{el},i} + d} = (1-\eta) \left(\frac{\mu_i^2}{\sigma_i^2 - \mu_i} \frac{1-\eta}{\eta} - 1 \right). \tag{57c}$$

One can show that the effective parameters $\sigma'_{u,i}$ and $\sigma'_{b,i}$ are always positive.

These parameter values can then be used to compute the probability distribution in the improved mean-field approximation (IMF) from the telegraph model,

$$P(n_i) = \frac{{\alpha'}_i^{n_i} e^{-\alpha'_i}}{n_i!} \frac{(\sigma'_{u,i})_{n_i}}{(\sigma'_{u,i} + \sigma'_{b,i})_{n_i}} \times M(\sigma'_{b,i}; \sigma'_{u,i} + \sigma'_{b,i} + n_i; \alpha'_i). \tag{58}$$

In Fig. 4 we compare the approximate distribution in Eq. (58) to the exact distribution obtained by stochastic simulations using the same model parameters α, σ_u and σ_b as in Fig. 3. We find an excellent agreement in all four cases, the constitutive, bursty, bimodal and nearly bimodal. We argue that option 2 performs significantly better than option 1 because the correlations between segments are solely due to promoter switching. These correlations can be better taken into account by adjusting the effective on and off rates, rather than keeping them the same.

Because the improved mean-field approximation matches only the first two moments, we wanted to check its accuracy in predicting the third standardised central moment or skewness. To this end we generated 8000 unique values of the parameters α, σ_u and σ_b. The values for each parameter were generated uniformly on a logarithmic scale using according to formula 2^k for integer k between -13 and 6. That gave us 20 values for each parameter spanning over five orders of magnitude in the range from $1.2 \cdot 10^{-4}$ to 64. For each combination we computed the exact skewness κ_i according to Eq. (38), and compared it to the one predicted by the improved mean-field approximation (computed from Eq. (58)),

$$\kappa_{1,i} = \frac{1}{\sigma_i^3} \Bigg(- \frac{2{\alpha'}_i^3 \eta(1-\eta)(2\eta-1)}{(1+\sigma'_{u,i}+\sigma'_{b,i})(2+\sigma'_{u,i}+\sigma'_{b,i})} + 3\sigma_i^2 - 2\mu_i \Bigg). \tag{59}$$

We did this for all segments on the gene, after which we selected the segment with the largest relative error $\epsilon = 100\% \cdot |\kappa_{1,i} - \kappa_i| / \kappa_i$. For this segment we further computed the RNAP number distribution in the improved mean-field approximation according to Eq. (58). Depending on the shape, we determined to which of the three categories the distribution corresponds to: unimodal (without an inflection point), bimodal and nearly bimodal (unimodal with an inflection point). Of 8000 distributions, 7190 were unimodal, 487 bimodal and 323 nearly bimodal.

The results comparing $\kappa_{1,i}$ versus exact κ_i are presented in Fig. 5. Due to the large range of κ_i, we separated the data according to $\kappa_i < 10$ (Fig. 5(a), linear scale) and $\kappa_i > 10$ (Fig. 5(b), log scale). Next, we inspected in more detail how the relative error ϵ is distributed among the four aforementioned categories. We focused on distributions for which the relative error $\epsilon > 10\%$. We found 74 such distributions in total, 64 were unimodal and 10 were bimodal. A closer inspection of these distributions revealed that they all had in common skewness $\kappa_i \approx 0$ which caused the relative error to be large. We inspected the distributions with the two largest ϵ values and used the Hellinger distance [26] to quantify their similarity compared to the exact distributions obtained by stochastic simulations. Indeed, we found very small Hellinger distances for both distributions, $3.7 \cdot 10^{-3}$ and $3.3 \cdot 10^{-3}$, confirming the high accuracy of our improved mean-field approximation.

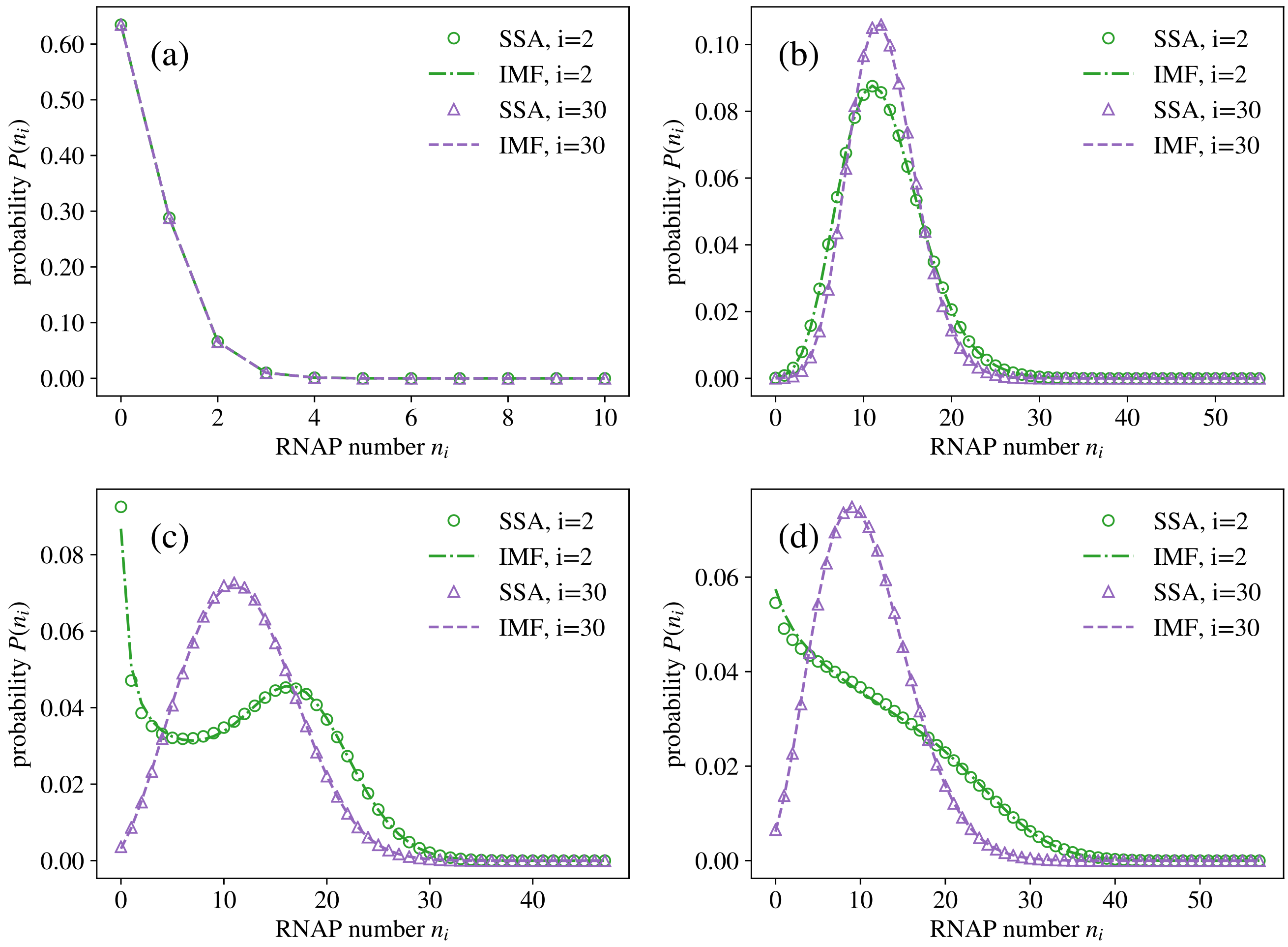

FIG. 4. RNAP number distribution in segments $i = 2$ and $i = L = 30$ in the improved mean-field approximation (IMF; the distribution is given by Eq. (58)), compared to the exact distribution obtained by the stochastic simulation algorithm (SSA). Model parameters are the same as in Fig. 3. The improved mean-field approximation performs significantly better than the naive mean-field approximation and is accurate even when the number correlations between segments are large. (a) Constitutive expression. (b) Bursty expression. (c) Bimodal distribution. (d) Nearly bimodal distribution.

C. Other applications: nuclear retention and export of mRNA

So far we have analysed the performance of the improved mean-field theory in the context of transcription elongation for which the assumption of uniform elongation seems reasonable. However, in other multistep downstream processes such as splicing or transport of nuclear mRNA to the cytoplasm, we cannot expect the rates of different downstream processing steps to be equal because each step may represent a physically different process.

Here we demonstrate the performance of the improved mean-field theory for the case of non-uniform rates. We consider the model for nuclear mRNA retention developed in Ref. [18]. A reaction scheme for this model with the corresponding rates is given by

$$G \underset{s_u}{\overset{s_b}{\rightleftharpoons}} G^*, \quad G \xrightarrow{r} G + P_1, \quad P_1 \xrightarrow{k_1} P_2 \xrightarrow{k_2} \emptyset. \qquad (60)$$

In this model, nuclear mRNA P_1 is produced at rate r from a gene in the on state and exported at rate k_1 from the nucleus to the cytoplasm (this is an $L = 2$ version of the reaction scheme (1)). The cytoplasmic mRNA P_2 is degraded at rate k_2. Parameters for this model were quantified experimentally for a set of genes in mouse liver cells in Ref. [18]. For our purposes we will use the following parameters obtained for Mlxipl gene,

$$r = 77\ \text{hr}^{-1},\ k_1 = 0.9\ \text{hr}^{-1},\ k_2 = 3.8\ \text{hr}^{-1}. \qquad (61)$$

In Ref. [18] only the fraction of active Mlxipl genes $\eta = 0.425$ was measured but not the absolute on and off rates s_u and s_b. In addition the coefficient of variance (CV) for cytoplasmic mRNA was measured to be

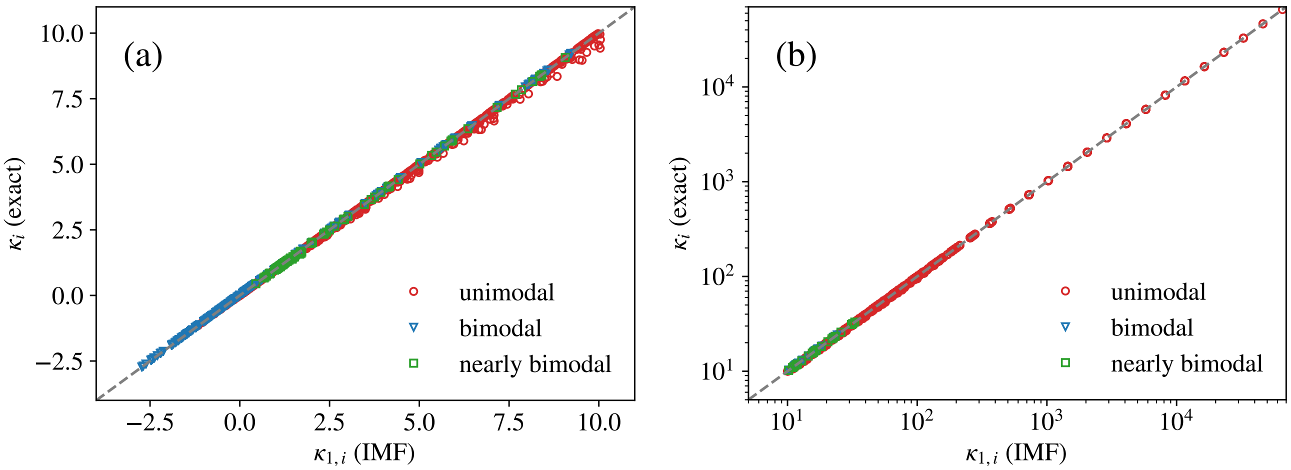

FIG. 5. Skewness $\kappa_{1,i}$ predicted by the improved mean-field approximation (IMF; see Eq. (59)) compared to the exact value κ_i (given by Eq. (38)) for 8000 combinations of α, σ_u and σ_b. Each point $(\kappa_{1,i}, \kappa_i)$ corresponds to one combination of parameters α, σ_u and σ_b. For each combination, the segment i with the largest relative error between $\kappa_{1,i}$ and κ_i was selected. The results were divided in two sets: (a) is for $\kappa_i < 10$ (linear scale) and (b) is for $\kappa_i > 10$ (log scale). The dashed line is for reference only and given by $\kappa_{1,i} = \kappa_i$.

CV$= 0.46$. By matching η and CV to those predicted by our theory (using the results for the mean and variance), we estimated the switching parameters: $s_u = 4.3$ hr^{-1} and $s_b = 5.7$ hr^{-1}. In addition to these values we considered two additional values of export rate k_1 that were two times smaller and larger than the experimental value.

Using these parameters we performed stochastic simulations of the model and compared the distributions of P_1 and P_2 with predictions of the improved mean-field theory. The results for nuclear mRNA are presented in Fig. 6(a) and for cytoplasmic mRNA in Fig. 6(b). In all cases we find excellent agreement with the predictions of the improved mean-field theory showing that it is equally accurate for non-equal rates of downstream processing steps.

D. Variation of the shape of the distribution with increasing number of downstream processing steps

In Fig. 4(c), we saw how a distribution that is bimodal for small i, can become unimodal for large enough i. The question we want to address here is whether this observation is a special case or if it generally holds.

Where the model is interpreted as one for RNAP dynamics during transcription, assuming uniform elongation rates across the gene ($k_{\mathrm{el,i}} = k_{\mathrm{el}}$) and no premature detachment ($d = 0$), using the exact Eqs. (22) and (31), one can easily deduce that the factor $\frac{\mu_i^2}{\sigma_i^2 - \mu_i}$ increases monotonically with i. Hence from Eqs. (57b) and (57c), it follows that the effective switching rates in the improved mean-field theory increase with distance i from the start site. This implies that any bimodality in the RNAP distribution is washed out as i increases because bimodality typically manifests due to slow switching between inactive and active states [27, 28].

A difficulty in generalizing this statement to the case of non-identical elongation rates and non-zero premature detachment is that in this case while the recurrence relations for the moments can be solved analytically, the solution is very complex for general number of downstream processing steps L. Instead it is easier to use Mathematica to solve the recurrence relations for $L = 1$, then for $L = 2$, etc and try to see a pattern in the algebraic equations. Using this method we verified that for any integer L, the factor $\frac{\mu_i^2}{\sigma_i^2 - \mu_i}$ increases monotonically with i. Hence by the same arguments as in the previous paragraph, we can generally state that for the general model (1), any bimodality in the distribution of P_i will decrease as i increases. Interpreting reaction scheme (1) as a model of the whole mRNA lifecycle, this implies that in a population of identical cells (since our model does not consider cell-to-cell variability), it is easier to observe bimodality in nascent/nuclear mRNA distributions than in cytoplasmic distributions.

The increase of the effective switching rates with i has also implications for the inference of the extent of transcriptional bursting [29]. This phenomenon is characterized by bursts of mRNA expression separated by long intervals of no expression, i.e. large mean burst size and low burst frequency. These two parameters are often estimated by fitting the telegraph model to measured distributions of the mRNA copy number [30]. Now it is well known that within the mathematical framework of the (non-leaky) telegraph model of gene expression, the mean burst size is the initiation rate divided by the off switching rate while the burst frequency is the on switch-

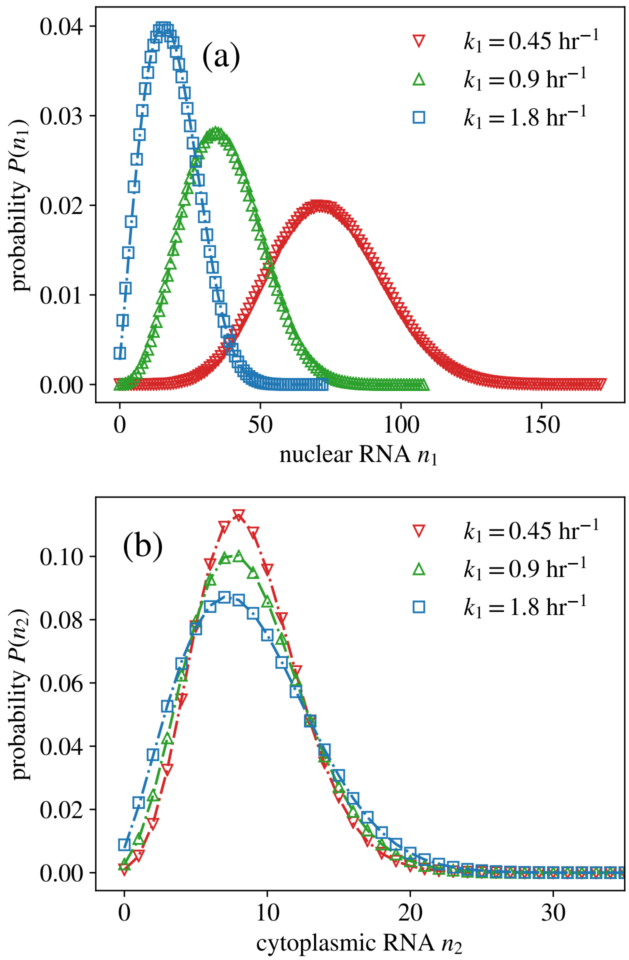

FIG. 6. Probability distributions of nuclear and cytoplasmic mRNA numbers in a two-state model of transcription with nuclear retention. Distributions were obtained by stochastic simulations (symbols) and compared to the prediction of the improved mean-field theory (lines). (a) Distribution of nuclear mRNA number $P(n_1)$. (b) Distribution of cytoplasmic mRNA number $P(n_2)$. The reaction scheme is given by (60) and the parameters are stated in the main text.

ing rate. We have shown above that the effective on rate increases with i which means that fitting the telegraph model to mRNA data primarily described by the i^{th} stage of the life cycle will necessarily overestimate the true burst frequency. The effective mean burst size is obtained by dividing Eq. (57a) by Eq. (57c). While the latter equation increases with i, the former equation can be independent or increase or decrease with i. Hence fitting the telegraph model to mRNA data primarily described by the i^{th} stage of the life cycle can under or overestimate the true burst size.

These results show that noise due to downstream processing steps can have a significant effect on the estimation of transcriptional parameters; this makes the case that for the accurate estimation of these parameters, the use of nascent mRNA data (which would correspond to small i in our model) is preferable to the use of cytoplasmic or whole cell mRNA data [31].

V. SUMMARY AND DISCUSSION

In this paper, we have devised a simple but very effective approximation to the steady-state distribution of the reaction scheme (1). The main idea behind our work was to approximate the dynamics of P_i in this model by that of an effective (non-leaky) telegraph model:

$$G \rightleftharpoons G^{\star}, \quad G \to G + P_i, \quad P_i \to \emptyset. \qquad (62)$$

In the improved mean-field approximation, the effective rates of the latter are deduced by matching the first two moments of P_i in the full and reduced models such that the ratio of the switching rates is the same in both models (but the absolute rates are not the same). This procedure is possible because the moments of both models are known exactly due to the linearity of their propensities. We have shown that this automatically means that the third-moments of both models are very close to each other, which suggests that the distributions of both models are also very close. By an extensive search over parameter space, we verified that the steady-state distribution of the reduced model distribution (which is known in closed-form) was very close to that obtained from stochastic simulations of the full model – the maximum Hellinger distance between the two distributions is of the order of 10^{-3} and in fact in all cases one could not easily distinguish by eye any difference between the two.

The study by Vastola et al [17] is the only other study of the steady-state distribution of (1) that we are aware of in the literature. In this paper, the same model is considered but there is no premature degradation and there is a non-zero transcription rate from the off state G^{*}. The model is studied in the context of transcription coupled to a switching gene with multiple downstream splicing steps. In this case, P_i can be interpreted as the number of mRNA molecules after $i - 1$ splicing steps. The paper use tools inspired by quantum mechanics, including ladder operators and Feynman-like diagrams, to write a formula for the joint distribution in terms of an infinite sum. This cannot be generally written in closed form. In particular the solution is in powers of the difference between the transcription rates of the active and inactive states, and since this is not typically small, one must compute a large number of series terms in order to closely approximate the correct answer. This is a disadvantage compared to our method which leads to a simple closed-form solution that is easy to evaluate and numerically stable. The disadvantage of our method relative to that of Vastola et al. is that it is not based on a systematic derivation; however as we have shown our method leads to an impressively accurate approximation of the full model over all parameter space. Another work related to our study is that by Bokes and Singh [32] which studied a one state gene system where mRNA molecules are produced in bursts (with a size that is distributed according to the geometric distribution) and then they are exported to the nucleus. This is likely [33] a special case of our model, i.e. the bursty limit where r and $s_b \to \infty$ at constant r/s_b for two segments $L = 2$. The authors find an exact solution to this special case; the marginal distribution of mRNA is found to be always unimodal. The advantage of our approximation is that it is valid not just in the bursty limit but all across parameter space, thus being able to capture the variation of the modality of the distribution through the mRNA lifecycle (all three types of distributions – unimodal, bimodal and nearly bimodal – have been measured [34–36]).

Finally we note that our present model could be improved to include further biological realism e.g. an arbitrary number of gene states and time-dependent switching rates. The latter is important since the presence of additional states is sometimes needed to explain the non-exponential duration of the off state for mammalian genes [37]; indeed a recent paper [38] found that a model equivalent to our reaction scheme (1) but with at most three gene states was sufficient to explain the general features of transcription and pervasive stochastic splice site selection. The extension of our model to consider time-dependent switching rates is also important to describe for example the activation of a promoter by a time-dependent transcriptional factor signal e.g. the identities and intensities of different stresses in budding yeast are transmitted by modulation of the amplitude, duration or frequency of nuclear translocation of the general stress response transcription factor Msn2 [39]. The modification of our novel mean-field approach to describe these more complex scenarios will be reported in a separate paper.

ACKNOWLEDGMENTS

We are grateful to Tatiana Filatova for critical reading of an earlier version of this manuscript. This work was supported by a Leverhulme Trust research award (RPG-2020-327).

[1] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins, Nature Reviews Genetics **6**, 451 (2005).

[2] D. Schnoerr, G. Sanguinetti, and R. Grima, Journal of Physics A: Mathematical and Theoretical **50**, 093001 (2017).

[3] C. W. Gardiner *et al.*, *Handbook of stochastic methods*, Vol. 3 (springer Berlin, 1985).

[4] T. Jahnke and W. Huisinga, Journal of mathematical biology **54**, 1 (2007).

[5] J. Peccoud and B. Ycart, Theoretical population biology **48**, 222 (1995).

[6] S. Iyer-Biswas, F. Hayot, and C. Jayaprakash, Physical Review E **79**, 031911 (2009).

[7] T. Zhou and J. Zhang, SIAM Journal on Applied Mathematics **72**, 789 (2012).

[8] L. Ham, D. Schnoerr, R. D. Brackston, and M. P. Stumpf, The Journal of Chemical Physics **152**, 144106 (2020).

[9] C. Shi, Y. Jiang, and T. Zhou, Biophysical Journal **119**, 1606 (2020).

[10] Z. Cao, T. Filatova, D. A. Oyarzún, and R. Grima, Biophysical Journal **119**, 1002 (2020).

[11] L. Ham, R. D. Brackston, and M. P. Stumpf, Physical review letters **124**, 108101 (2020).

[12] J. Dattani and M. Barahona, Journal of The Royal Society Interface **14**, 20160833 (2017).

[13] Z. Cao and R. Grima, Proceedings of the National Academy of Sciences **117**, 4682 (2020).

[14] R. Perez-Carrasco, C. Beentjes, and R. Grima, Journal of the Royal Society Interface **17**, 20200360 (2020).

[15] C. Zhu, G. Han, and F. Jiao, Complexity **2020** (2020).

[16] T. Filatova, N. Popovic, and R. Grima, Bulletin of Mathematical Biology **83**, 1 (2021).

[17] J. J. Vastola, G. Gorin, L. Pachter, and W. R. Holmes, arXiv preprint arXiv:2103.10992 (2021).

[18] K. Bahar Halpern, I. Caspi, D. Lemze, M. Levy, S. Landen, E. Elinav, I. Ulitsky, and S. Itzkovitz, Cell Reports **13**, 2653 (2015).

[19] D. Cao and R. Parker, Rna **7**, 1192 (2001).

[20] M. Z. Ali, S. Choubey, D. Das, and R. C. Brewster, Biophysical Journal **118**, 1769 (2020).

[21] N. Van Kampen, in *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 2007) 3rd ed., p. 139.

[22] M. Abramowitz, I. A. Stegun, and R. H. Romer, Handbook of mathematical functions with formulas, graphs, and mathematical tables (1988).

[23] J. M. Pedraza and J. Paulsson, Science **319**, 339 (2008).

[24] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, *et al.*, Nature **560**, 494 (2018).

[25] We note that in the derivation of cov(n_i, n_j) we have used the fact that $\gamma_{i,j}(0,0) = 1$, which is not at all obvious that is true. One can prove this relation by showing that $\gamma_{i,j}(0,0) = \gamma_{i,j-1}(0,0)/2 + \gamma_{i-1,j}(0,0)/2$ for any $i, j \geq 2$, and also that $\gamma_{i,1}(0,0) = \gamma_{1,j}(0,0) = 1$ for any $i, j \geq 1$. This recurrence relation has a simple solution $\gamma_{ij}(0,0) = 1$ for any $i, j \geq 1$.

[26] The Hellinger distance takes values between 0 and 1, where 0 is achieved if two distributions are equal, and 1 is achieved if one distribution assigns probability 0 to every set to which the other assigns a positive probability and vice versa.

[27] P. Thomas, N. Popović, and R. Grima, Proceedings of the National Academy of Sciences **111**, 6994 (2014).

[28] T. Plesa, R. Erban, and H. G. Othmer, European Journal of Applied Mathematics **30**, 887 (2019).

[29] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger, Proceedings of the National Academy of Sciences **109**, 17454 (2012).

[30] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg, Nature **565**, 251 (2019).

[31] H. Xu, S. O. Skinner, A. M. Sokac, and I. Golding, Physical review letters **117**, 128101 (2016).

[32] A. Singh and P. Bokes, Biophysical journal **103**, 1087 (2012).

[33] It is yet to be formally proven using singular perturbation theory that this is indeed the case. A proof can be constructed similar to that reported for other systems [13].

[34] Z. S. Singer, J. Yong, J. Tischler, J. A. Hackett, A. Altinok, M. A. Surani, L. Cai, and M. B. Elowitz, Molecular cell **55**, 319 (2014).

[35] H. Xu, L. A. Sepúlveda, L. Figard, A. M. Sokac, and I. Golding, Nature methods **12**, 739 (2015).

[36] S. Hocine, M. Vera, D. Zenklusen, and R. H. Singer, Scientific reports **5**, 1 (2015).

[37] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Science **332**, 472 (2011).

[38] Y. Wan, D. G. Anastasakis, J. Rodriguez, M. Palangat, P. Gudla, G. Zaki, M. Tandon, G. Pegoraro, C. C. Chow, M. Hafner, *et al.*, Cell **184**, 2878 (2021).

[39] N. Hao and E. K. O'Shea, Nature structural & molecular biology **19**, 31 (2012).