Approximate probability distributions of the master equation

Philipp Thomas[*]

School of Mathematics and School of Biological Sciences, University of Edinburgh, Edinburgh EH8 9YL, United Kingdom

Ramon Grima[†]

School of Biological Sciences, University of Edinburgh, Edinburgh EH8 9YL, United Kingdom

Master equations are common descriptions of mesoscopic systems. Analytical solutions to these equations can rarely be obtained. We here derive an analytical approximation of the time-dependent probability distribution of the master equation using orthogonal polynomials. The solution is given in two alternative formulations: a series with continuous and a series with discrete support, both of which can be systematically truncated. While both approximations satisfy the system size expansion of the master equation, the continuous distribution approximations become increasingly negative and tend to oscillations with increasing truncation order. In contrast, the discrete approximations rapidly converge to the underlying non-Gaussian distributions. The theory is shown to lead to particularly simple analytical expressions for the probability distributions of molecule numbers in metabolic reactions and gene expression systems.

PACS number(s): 02.50.Ey, 87.18.Tt, 05.40.−a, 87.23.−n

I. INTRODUCTION

Master equations are commonly used to describe fluctuations of particulate systems. In most instances, however, the number of reachable states is so large that their combinatorial complexity prevents one from obtaining analytical solutions to these equations. Explicit solutions are known only for certain classes of linear birth-death processes [1], under detailed balance conditions [2], or for particularly simple examples in stationary conditions [3]. Considerable effort has been undertaken to approximate the solution of the master equation under more general conditions, including time dependence and conditions lacking detailed balance [4,5].

A common technique addressing this issue was given by van Kampen in terms of the system size expansion (SSE) [6]. The method assumes the existence of a specific parameter, termed the system size, for which the master equation approaches a deterministic limit as its value is taken to infinity. The leading-order term of this expansion describes small fluctuations about this limit in terms of a Gaussian probability density, called the linear noise approximation (LNA). This approximation has been widely applied in biochemical kinetics [7] but also in the theory of polymer assembly [8], epidemics [9], economics [10], and machine learning [11]. The benefit of the LNA is that it yields generic expressions for the probability density. Its deficiency lies in the fact that, strictly speaking, it is valid only in the limit of infinite system size. Hence one generally suspects that its predictions become inaccurate when one studies fluctuations that are not too small compared to the mean and therefore implying non-Gaussian statistics.

Higher-order terms in the SSE have been employed to calculate non-Gaussian corrections to the LNA for the first few moments [12–14]; alternative methods are based on moment closure [15]. It is, however, the case that the knowledge of a limited number of moments does not allow us to uniquely determine the underlying distribution functions. Reconstruction of the probability distribution therefore requires additional approximations such as the maximum entropy principle [16] or the truncated moment generating function [17], which generally yield different results. While the accuracy of these repeated approximations remains unknown, analytical expressions for the probability density can rarely be obtained or might not even exist [18]. A systematic investigation of the distributions implied by the higher-order terms in the SSE, without resorting to moments, is therefore still missing.

We here analytically derive, for the first time, as far as we know, a closed-form series expansion of the probability distribution underlying the master equation. We proceed by outlining the expansion of the master equation in Sec. I and briefly review the solution of the leading-order terms given by the LNA in Sec. II. While commonly the SSE is truncated at this point, we show that the higher-order terms can be obtained using an asymptotic expansion of the continuous probability density. The resulting series is given in terms of orthogonal polynomials and can be truncated systematically to any desired order in the inverse system size. Analytical expressions are given for the expansion coefficients.

Thereby we establish two alternative formulations of this expansion: a continuous and a discrete one, both satisfying the expansion of the master equation. We show that for linear birth-death processes, the continuous approximation often fails to converge reasonably fast. In contrast, the discrete approximation introduced in Sec. III accurately converges to the true distribution with increasing truncation order. In Sec. IV, we show that for nonlinear birth-death processes, renormalization is required for achieving rapid convergence of the series. Our analysis is motivated by the use of simple examples throughout. Using a common model of gene expression, we conclude in Sec. VI that the new method allows us to predict the full time dependence of the molecule number distribution.

————————
[*]philipp.thomas@ed.ac.uk
[†]ramon.grima@ed.ac.uk

II. SYSTEM SIZE EXPANSION

As a starting point, we focus on the master equation formulation of biochemical kinetics. We therefore consider a set of R chemical reactions involving a single species confined in a well-mixed volume Ω. Note that for chemical systems the system size coincides with the reaction volume. We denote by S_r the net change in the molecule numbers in the rth reaction and by $\gamma_r(n,\Omega)$ the probability per unit time for this reaction to occur. The probability of finding n molecules in the volume Ω at time t, denoted by $P(n,t)$, then obeys the master equation

$$\frac{dP(n,t)}{dt} = \sum_{r=1}^{R}(E^{-S_r}-1)\gamma_r(n,\Omega)P(n,t), \qquad (1)$$

where E^{-S_r} is the step operator defined as $E^{-S_r}g(n) = g(n-S_r)$ for any function $g(n)$ of the molecule numbers [6]. Note that throughout the article, deterministic initial conditions are assumed. The system size expansion now proceeds by separating the instantaneous concentration into a deterministic part, given by the solution of the rate equations $[X]$, and a fluctuating part ϵ,

$$\frac{n}{\Omega} = [X] + \Omega^{-1/2}\epsilon, \qquad (2)$$

which is van Kampen's ansatz. The expansion of the master equation can be summarized in three steps:

(i) Using Eq. (2) one expands the step operator

$$\begin{aligned}
E^{-S_r}\gamma_r(n,\Omega)P(n,t) &= \gamma_r(n-S_r,\Omega)P(n-S_r,t)\\
&= e^{-\Omega^{-1/2}S_r\partial_\epsilon}\gamma_r(\Omega[X]+\Omega^{1/2}\epsilon,\Omega)\\
&\quad \times P(\Omega[X]+\Omega^{1/2}\epsilon,t), \qquad (3)
\end{aligned}$$

where ∂_ϵ denotes $\frac{\partial}{\partial\epsilon}$.

(ii) Next, the probability for the molecule numbers is cast into a probability density $\Pi(\epsilon,t)$ for the fluctuations using van Kampen's ansatz,

$$\Pi(\epsilon,t) = \Omega^{1/2}P(\Omega[X]+\Omega^{1/2}\epsilon,t), \qquad (4)$$

which is essentially a change of variables. Note that this step implicitly assumes a continuous approximation $\Pi(\epsilon,t)$ of the probability distribution as thought in the original derivation of van Kampen [6].

(iii) It remains to expand the propensity about the deterministic limit

$$\gamma_r(\Omega[X]+\Omega^{1/2}\epsilon,\Omega) = \sum_{k=0}^{\infty}\Omega^{-k/2}\frac{\epsilon^k}{k!}\frac{\partial^k\gamma_r(\Omega[X],\Omega)}{\partial[X]^k}. \qquad (5)$$

Note that $\gamma_r(\Omega[X],\Omega)$ is just the propensity evaluated at the macroscopic concentration and hence it must depend explicitly on Ω. We assume that the propensity possesses a power series in the inverse volume

$$\gamma_r(\Omega[X],\Omega) = \Omega\sum_{s=0}^{\infty}\Omega^{-s}f_r^{(s)}([X]). \qquad (6)$$

For mass-action kinetics, for instance, the propensity is given by $\gamma_r(n,\Omega) = \Omega^{1-\ell_r}k_r\ell_r!\binom{n}{\ell_r}$, where ℓ_r is the reaction order of the rth reaction. Using the Taylor expansion of the binomial coefficient, we have $f_r^{(0)}([X]) = k_r[X]^{\ell_r}$,

$f_r^{(s)}([X]) = k_r[X]^{\ell_r-s}\mathcal{S}_{\ell_r,\ell_r-s}$, and $f_r^{(s)} = 0$ for $s \geqslant \ell_r$, where $\mathcal{S}$ denotes the Stirling numbers of the first kind. Note also that effective propensities being deduced from mass action kinetics have an expansion similar to Eq. (6). The Michaelis-Menten propensity $\gamma_r(n,\Omega) = \Omega k_r\frac{n}{n+K\Omega}$ [19], for instance, has $f_r^{(0)}([X]) = k_r\frac{[X]}{[X]+K}$ and $f_r^{(s)}([X]) = 0$ for $s > 0$.

Substituting now Eqs. (3)–(6) into Eq. (1) and rearranging the result in powers of $\Omega^{-1/2}$, we find

$$\begin{aligned}
&\left(\frac{\partial}{\partial t} - \Omega^{1/2}\frac{d[X]}{dt}\frac{\partial}{\partial\epsilon}\right)\Pi(\epsilon,t)\\
&= \left[-\Omega^{1/2}\sum_{r=1}^{R}S_r f_r^{(0)}([X])\frac{\partial}{\partial\epsilon} + \sum_{k=0}^{N}\Omega^{-k/2}\mathcal{L}_k\right]\Pi(\epsilon,t)\\
&\quad + O[\Omega^{-(N+1)/2}]. \qquad (7)
\end{aligned}$$

Equating terms to order $\Omega^{1/2}$ yields the macroscopic rate equation

$$\frac{d[X]}{dt} = \sum_{r=1}^{R}S_r f_r^{(0)}([X]). \qquad (8)$$

The higher-order terms in the expansion of the master equation can be written out explicitly,

$$\mathcal{L}_k = \sum_{s=0}^{\lceil k/2\rceil}\sum_{p=1}^{k-2(s-1)}\frac{\mathcal{D}_{p,s}^{k-p-2(s-1)}}{p![k-p-2(s-1)]!}(-\partial_\epsilon)^p\epsilon^{k-p-2(s-1)}, \qquad (9)$$

where $\lceil\cdot\rceil$ denotes the ceiling value and the coefficients are given by

$$\mathcal{D}_{p,s}^q = \sum_{r=1}^{R}(S_r)^p\partial_{[X]}^q f_r^{(s)}([X]), \qquad (10)$$

which depend explicitly on the solution of the rate equation (8). Note that in the following the abbreviation $\mathcal{D}_p^q = \mathcal{D}_{p,0}^q$ is used.

III. EXPANSION OF THE CONTINUOUS PROBABILITY DENSITY

We here study the time-dependent solution of the partial differential equation approximation of the master equation, Eq. (7). We therefore expand the probability density of Eq. (4),

$$\Pi(\epsilon,t) = \sum_{j=0}^{N}\Omega^{-j/2}\pi_j(\epsilon,t) + O[\Omega^{-(N+1)/2}], \qquad (11)$$

which also allows the expansion of the time-dependent moments to be deduced in closed-form. Finally, we recover the stationary solution as a particular case.

A. Linear noise approximation

Substituting Eq. (11) into Eq. (7) and equating terms to order $O(\Omega^0)$, we find

$$\left(\frac{\partial}{\partial t} - \mathcal{L}_0\right)\pi_0 = 0, \qquad (12)$$

where $\mathcal{L}_0 = -\partial_\epsilon \mathcal{J}\epsilon + \frac{1}{2}\partial_\epsilon^2 \mathcal{D}_2^0$ is a Fokker-Planck operator with linear coefficients and $\mathcal{J} = \mathcal{D}_1^1$ is the Jacobian of the rate equation. The probability density of fluctuations about the macroscopic concentration, described by ϵ, is given by a centered Gaussian,

$$\pi_0(\epsilon, t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} \exp\left[-\frac{\epsilon^2}{2\sigma^2(t)}\right], \qquad (13)$$

which acquires time dependence via its variance $\sigma^2(t)$. The latter satisfies

$$\frac{\partial\sigma^2}{\partial t} = 2\mathcal{J}(t)\sigma^2 + \mathcal{D}_2^0(t), \qquad (14)$$

which is the familiar LNA result [6]. In the following we will drop the time dependence of the coefficients for convenience of notation.

B. Higher-order terms

Substituting Eq. (11) into Eq. (7), rearranging the remaining terms, and equating terms to order $\Omega^{-j/2}$, we find

$$\left(\frac{\partial}{\partial t} - \mathcal{L}_0\right)\pi_j(\epsilon, t) = \mathcal{L}_1\pi_{j-1} + \cdots + \mathcal{L}_j\pi_0$$

$$= \sum_{k=1}^{j} \mathcal{L}_k\pi_{j-k}(\epsilon, t). \qquad (15)$$

This system of partial differential equations can be solved using the eigenfunction approach. We consider

$$\left(\frac{\partial}{\partial t} - \mathcal{L}_0\right)\Psi_m = \lambda_m\Psi_m, \qquad (16)$$

which is solved by $\lambda_m = -m\mathcal{J}$ and $\Psi_m = \psi_m(\epsilon, t)\pi_0(\epsilon, t)$ with

$$\psi_m(\epsilon, t) = \pi_0^{-1}(-\partial_\epsilon)^m\pi_0 = \frac{1}{\sigma^m}H_m\left(\frac{\epsilon}{\sigma}\right). \qquad (17)$$

The functions H_m denote the Hermite orthogonal polynomials which are given explicitly in Appendix A. To verify the solution of the eigenvalue problem, we set $\Psi_{m+1} = (-\partial_\epsilon)\Psi_m$ and observe that $(\partial_t - \mathcal{L}_0)\Psi_{m+1} = -\mathcal{J}\Psi_{m+1} - \partial_\epsilon(\partial_t - \mathcal{L}_0)\Psi_m$. Using this in Eq. (16), we obtain $\lambda_{m+1} = (-\mathcal{J} + \lambda_m)$ from which the result follows because $\lambda_0 = 0$ and $\Psi_0 = \pi_0$.

Using the completeness of the eigenfunctions, we can write $\pi_j(\epsilon, t) = \sum_{m=0}^{\infty} a_m^{(j)}(t)\psi_m(\epsilon, t)\pi_0(\epsilon, t)$. We verify in Appendix B that the jth order term in the expansion involves only the first $N_j = 3j$ eigenfunctions. The continuous SSE approximation is consequently given by the asymptotic expansion

$$\Pi(\epsilon, t) = \pi_0(\epsilon, t)\left[1 + \sum_{j=1}^{N} \Omega^{-j/2} \sum_{m=1}^{N_j} a_m^{(j)}(t)\psi_m(\epsilon, t)\right]$$

$$+ O[\Omega^{-(N+1)/2}], \qquad (18)$$

for which the coefficients can be determined using the orthogonality of the functions ψ_m, i.e., $\frac{\sigma^{2n}}{n!}\int d\epsilon\,\psi_n(\epsilon, t)\psi_m(\epsilon, t)\pi_0(\epsilon, t) = \delta_{m,n}$.

C. The equation for the expansion coefficients

The coefficients $a_n^{(j)}$ are now determined by inserting the expansion of π_j into Eq. (15), multiplying the result by $\frac{\sigma^{2n}}{n!}\int d\epsilon\,\psi_n(\epsilon, t)$, and performing the integration. Using Eq. (16), the left-hand side of Eq. (15) becomes

$$\frac{\sigma^{2n}}{n!}\sum_m \int d\epsilon\,\psi_n(\epsilon, t)\left(\frac{\partial}{\partial t} - \mathcal{L}_0\right)a_m^{(j)}\psi_m(\epsilon, t)\pi_0(\epsilon, t)$$

$$= \left(\frac{\partial}{\partial t} - n\mathcal{J}\right)a_n^{(j)}. \qquad (19)$$

The calculation of terms in the summation on the right-hand side of Eq. (15) is greatly simplified by defining the integral

$$\mathcal{I}_{mn}^{\alpha\beta} = \frac{\sigma^{2n}}{n!\alpha!\beta!}\int d\epsilon\,\psi_n(\epsilon, t)(-\partial_\epsilon)^\alpha\epsilon^\beta\psi_m(\epsilon, t)\pi_0(\epsilon, t), \quad (20)$$

which yields

$$\frac{\sigma^{2n}}{n!}\int d\epsilon\,\psi_n(\epsilon, t)\mathcal{L}_k\psi_m(\epsilon, t)\pi_0(\epsilon, t)$$

$$= \sum_{s=0}^{\lceil k/2 \rceil}\sum_{p=1}^{k-2(s-1)} \mathcal{D}_{p,s}^{k-p-2(s-1)}\mathcal{I}_{mn}^{p,k-p-2(s-1)}. \qquad (21)$$

Using Eqs. (19) and (21) in Eq. (15), we find that the coefficients satisfy the following set of ordinary differential equations:

$$\left(\frac{\partial}{\partial t} - n\mathcal{J}\right)a_n^{(j)}$$

$$= \sum_{k=1}^{j}\sum_{m=0}^{N_{j-k}} a_m^{(j-k)}\sum_{s=0}^{\lceil k/2 \rceil}\sum_{p=1}^{k-2(s-1)} \mathcal{D}_{p,s}^{k-p-2(s-1)}\mathcal{I}_{mn}^{p,k-p-2(s-1)},$$

$$\qquad (22)$$

where we have assumed $a_n^{(j)} = 0$ for $n > N_j$. Explicitly, the nonzero integrals are given by

$$\mathcal{I}_{mn}^{\alpha\beta} = \frac{\sigma^{\beta-\alpha+n-m}}{\alpha!}\sum_{s=0}^{\min(n-\alpha,m)}\binom{m}{s}$$

$$\times \frac{[\beta + \alpha + 2s - (m+n) - 1]!!}{[\beta + \alpha + 2s - (m+n)]!(n - \alpha - s)!}, \qquad (23)$$

and zero for odd $(\alpha + \beta) - (m+n)$. Here $(2k-1)!! = \frac{(2k)!}{2^k k!}$ is the double factorial. Along with Eq. (23), in Appendix B we verify the following two important properties of the asymptotic series solution given deterministic initial conditions: (i) We have $N_j = 3j$ and hence Eq. (22) indeed yields a finite number of equations and (ii) $a_n^{(j)}$ vanishes for all times when $(n + j)$ is odd.

Finally, we note that $\mathcal{D}_{p,s}^q$ and $\mathcal{I}_{mn}^{pq}$ are generally time dependent because they are functions of the solution of the rate equation and the LNA variance. Explicit expressions for the approximate probability density can now be evaluated to any desired order.

D. Moments of the distribution

The solution for the probability density enables one to derive closed-form expressions for the moments. These are

obtained by multiplying Eq. (18) by $\int d\epsilon\, \epsilon^\beta$ and performing the integration using Eq. (B4) of Appendix B. We find

$$\langle \epsilon^\beta \rangle = \sum_{j=0}^{N} \Omega^{-j/2} \sum_{k=0}^{\lfloor \beta/2 \rfloor} \frac{\beta!}{2^k k!} \sigma^{2k} a_{\beta-2k}^{(j)} + O[\Omega^{-(N+1)/2}], \quad (24)$$

where $a_0^{(j)} = \delta_{0,j}$ and $\lfloor \cdot \rfloor$ denotes the floor value. In particular, it follows that the mean and variance are given by $\langle \epsilon \rangle = \sum_{j=1}^{N} \Omega^{-j/2} a_1^{(j)} + O[\Omega^{-(N+1)/2}]$ and $\langle \epsilon^2 \rangle = \sigma^2 + 2\sum_{j=1}^{N} \Omega^{-j/2} a_2^{(j)} + O[\Omega^{-(N+1)/2}]$.

It is now evident that the coefficients of the expansion are intricately related to the system size expansion of the distribution moments. Naturally, one may seek to invert this relation. Indeed, as we show in Appendix C, given the expansion for a finite set moments, the coefficients in Eq. (18) can be uniquely determined. In particular, to construct the probability density to order $\Omega^{-j/2}$ one requires the expansion of the first $3j$ moments to the same order. Thus the problem of moments provides an equivalent route of systematically constructing solutions to the master equation.

E. Solution in stationary conditions

Of particular interest is the expansion of the probability density under stationary conditions. Implicitly, we assume here that the rate equation, Eq. (8), has a single asymptotically stable fixed point, and hence the LNA variance is given by $\sigma^2 = \mathcal{D}_2^0 / (-2\mathcal{J})$. Setting the time derivative on the left-hand side of Eq. (22) to zero, we find that the coefficients of Eq. (18) can be expressed in terms of lower-order ones,

$$a_n^{(j)} = -\frac{1}{n\mathcal{J}} \sum_{k=1}^{j} \sum_{s=0}^{\lceil k/2 \rceil} \sum_{p=1}^{k-2(s-1)}$$
$$\times \mathcal{D}_{p,s}^{k-p-2(s-1)} \sum_{m=0}^{3(j-k)} a_m^{(j-k)} \mathcal{I}_{mn}^{p,k-p-2(s-1)}. \quad (25)$$

For example, truncating after terms of order Ω^{-1}, we obtain

$$\Pi(\epsilon) = \pi_0(\epsilon) + \Omega^{-1/2} \big[a_1^{(1)} \psi_1(\epsilon) + a_3^{(1)} \psi_3(\epsilon) \big] \pi_0(\epsilon)$$
$$+ \Omega^{-1} \big[a_2^{(2)} \psi_2(\epsilon) + a_4^{(2)} \psi_4(\epsilon) + a_6^{(2)} \psi_6(\epsilon) \big] \pi_0(\epsilon)$$
$$+ O(\Omega^{-3/2}). \quad (26)$$

The nonzero coefficients to order $\Omega^{-1/2}$ are given by

$$a_1^{(1)} = -\frac{\sigma^2 \mathcal{D}_1^2}{2\mathcal{J}} - \frac{\mathcal{D}_{1,1}^0}{\mathcal{J}},$$
$$a_3^{(1)} = -\frac{\sigma^4 \mathcal{D}_1^2}{6\mathcal{J}} - \frac{\sigma^2 \mathcal{D}_2^1}{6\mathcal{J}} - \frac{\mathcal{D}_3^0}{18\mathcal{J}}, \quad (27)$$

while those to order Ω^{-1} are

$$a_2^{(2)} = -a_1^{(1)}\left(\frac{\mathcal{D}_{1,1}^0}{2\mathcal{J}} + \frac{\mathcal{D}_2^1}{4\mathcal{J}} + \frac{3\sigma^2 \mathcal{D}_1^2}{4\mathcal{J}} \right) - a_3^{(1)} \frac{3\mathcal{D}_1^2}{2\mathcal{J}}$$
$$- \frac{\mathcal{D}_{2,1}^0}{4\mathcal{J}} - \frac{\sigma^2 \mathcal{D}_{1,1}^1}{2\mathcal{J}} - \frac{\sigma^2 \mathcal{D}_2^2}{8\mathcal{J}} - \frac{\sigma^4 \mathcal{D}_1^3}{4\mathcal{J}},$$
$$a_4^{(2)} = -a_1^{(1)}\left(\frac{\mathcal{D}_3^0}{24\mathcal{J}} + \frac{\sigma^2 \mathcal{D}_2^1}{8\mathcal{J}} + \frac{\sigma^4 \mathcal{D}_1^2}{8\mathcal{J}} \right) - \frac{\mathcal{D}_4^0}{96\mathcal{J}} - \frac{\sigma^2 \mathcal{D}_3^1}{24\mathcal{J}}$$

$$- \frac{\sigma^4 \mathcal{D}_2^2}{16\mathcal{J}} - \frac{\sigma^6 \mathcal{D}_1^3}{24\mathcal{J}} - a_3^{(1)}\left(\frac{\mathcal{D}_{1,1}^0}{4\mathcal{J}} + \frac{3\mathcal{D}_2^1}{8\mathcal{J}} + \frac{7\sigma^2 \mathcal{D}_1^2}{8\mathcal{J}} \right),$$
$$a_6^{(2)} = \frac{1}{2}\big(a_3^{(1)} \big)^2. \quad (28)$$

The accuracy of this distribution approximation is studied through an example in the following.

F. The continuous approximation fails under low molecule number conditions

We now study the SSE solution for a linear birth-death process, i.e., its propensities depend at most linearly on the molecular populations. Specifically, we consider the synthesis and decay of a molecular species X,

$$\varnothing \underset{k_1}{\overset{k_0}{\rightleftharpoons}} X. \quad (29)$$

The master equation is constructed using $S_1 = +1$, $\gamma_1 = \Omega k_0$, $S_2 = -1$, $\gamma_2 = k_1 n$, and $R = 2$ in Eq. (1). The exact stationary solution of the master equation is a Poisson distribution with mean $\Omega[X]$, where $[X] = k_0/k_1$. The coefficients in Eq. (10) are then given by

$$\mathcal{D}_n^m = \delta_{m,0} k_0 + (-1)^n k_1 (\delta_{m,0}[X] + \delta_{m,1}), \quad (30)$$

and $\mathcal{D}_{n,s}^m = 0$ for $s > 0$. The leading-order corrections to the LNA given by Eqs. (26)–(28) lead to very compact expressions for the expansion coefficients and are given by

$$a_3^{(1)} = \frac{[X]}{6}, \quad a_4^{(2)} = \frac{[X]}{24}, \quad a_6^{(2)} = \frac{[X]^2}{72} \quad (31)$$

and $a_1^{(1)} = a_2^{(2)} = 0$.

Though the continuous approximation is expected to perform well at large values of Ω, we are particularly interested in its performance when the value of Ω is decreased. Since the expansion is carried out at constant average concentration, low values of Ω typically imply low numbers of molecules and non-Gaussian distributions. In Fig. 1(a) we show that for parameters yielding half a molecule on average, the continuous approximation obtained in this section, given by Eq. (18) together with Eqs. (25) and (30), is unsatisfactory since as higher orders are taken into account, one observes large oscillations in the tails of the distribution. In the following section we show that the disagreement arises due to the assumption that the support of the distribution is continuous rather than discrete as implied by the master equation.

IV. DISCRETE APPROXIMATION OF THE PROBABILITY DISTRIBUTION

The aim of this paragraph is to establish a discrete formulation of the distribution approximations. To clarify this issue, we note that the exact characteristic function $G(k,t) = \sum_{n=0}^{\infty} e^{ikn} P(n,t)$ is a 2π-periodic function and hence can be inverted as follows:

$$P(n,t) = \int_{-\pi}^{\pi} \frac{dk}{2\pi} e^{-ikn} G(k,t). \quad (32)$$

We now associate our continuous approximation, Eq. (18), with this characteristic function, i.e., $G(k,t) = \int_{-\infty}^{\infty} d\epsilon\, e^{ik\Omega[X] + \Omega^{-1/2}\epsilon} \Pi(\epsilon,t)$. Substituting this
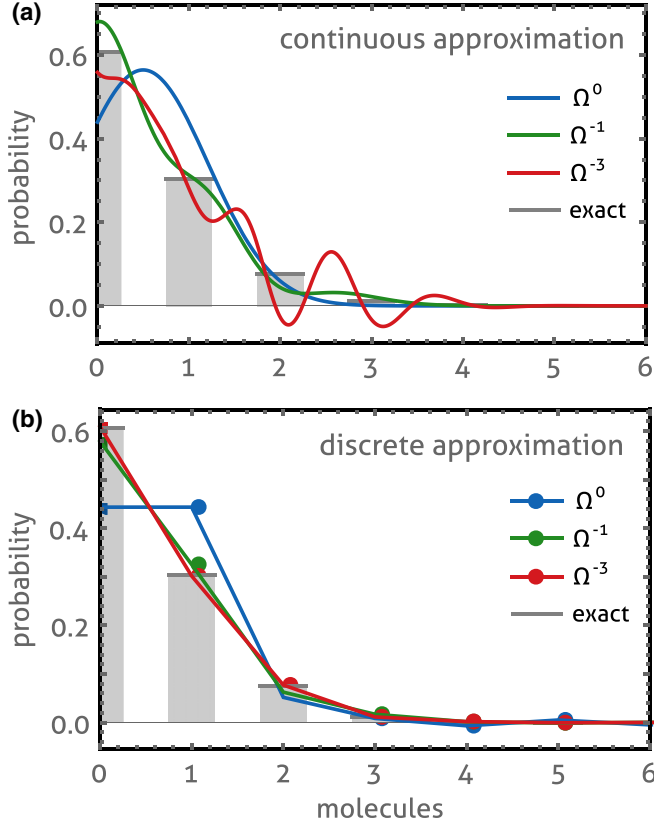
FIG. 1. (Color online) Linear birth-death process. We consider the reaction system (29) in stationary conditions. (a) We compare the exact Poisson distribution (gray) to the continuous SSE approximation [Eq. (18) together with Eqs. (25) and (30)] truncated after the Ω^0 (LNA, blue line), Ω^{-1} (green), and Ω^{-3} terms (red) for parameter values $k_0 = 0.5, k_1 = 1$, and $\Omega = 1$ giving half a molecule on average. We observe that the continuous approximation becomes increasingly negative and tends to oscillations with increasing truncation order. (b) In contrast, the discrete approximation shows no oscillations, and the overall agreement with the exact Poisson distribution (gray bars) improves with increasing truncation order.

together with Eq. (11) into Eq. (32) one establishes a connection formula between these discrete and continuous approximations via the convolution

$$P(n,t) = \sum_{j=0}^{N} \Omega^{-j/2} \int_{-\infty}^{\infty} d\epsilon \, K(n - \Omega[X] - \Omega^{1/2}\epsilon)\pi_j(\epsilon,t)$$
$$+ \, O(\Omega^{-(N+1)/2}), \tag{33}$$

with kernel

$$K(s) = \int_{-\pi}^{\pi} \frac{dk}{2\pi} e^{-iks} = \frac{\sin(\pi s)}{\pi s}. \tag{}$$

The convolution can be used to define the derivatives of the discrete probability via

$$\partial_n P(n,t) = \int_{-\infty}^{\infty} d\epsilon \, K(n - \Omega[X] - \Omega^{1/2}\epsilon)(\Omega^{-1/2}\partial_\epsilon)\Pi(\epsilon,t), \tag{34}$$

and hence it satisfies $E^{-S_j} P(n,t) = \int_{-\infty}^{\infty} d\epsilon \, K(n - \Omega[X] - \Omega^{1/2}\epsilon)e^{-\Omega^{-1/2}\partial_\epsilon S_j} \Pi(\epsilon,t)$, as well as $\gamma_j(n,\Omega)P(n,t) = \int_{-\infty}^{\infty} d\epsilon \, K(n - \Omega[X] - \Omega^{1/2}\epsilon)\gamma_j(\Omega[X] + \Omega^{1/2}\epsilon,\Omega)\Pi(\epsilon,t)$ for analytic γ_j. It then follows from the fact that $P(n,t)$ and $\Omega^{1/2}\Pi(\Omega^{-1/2}(n - \Omega[X]),t)$ have the same characteristic function expansion that (i) both approximations possess the same asymptotic expansion of their moments and that (ii) they satisfy the same expansion of the master equation.

For example, to leading order Ω^0, Eq. (33) replaces the conventional continuous LNA estimate, π_0 given by Eq. (13), with a discrete approximation,

$$P_0(n,t) = \frac{1}{2} \frac{e^{-\frac{y^2}{2\Sigma^2}}}{\sqrt{2\pi}\Sigma} \left[\text{erf}\left(\frac{iy + \pi\Sigma^2}{\sqrt{2}\Sigma}\right) - \text{erf}\left(\frac{iy - \pi\Sigma^2}{\sqrt{2}\Sigma}\right) \right], \tag{35}$$

where $y = n - \Omega[X]$, $\Sigma^2 = \Omega\sigma^2$ is the LNA's estimate for the variance of molecule numbers, and erf is the error function defined by $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$.

Associating the $\Omega^{-j/2}$ term of Eq. (18) with π_j in Eq. (33), higher-order approximations can now be obtained from

$$P(n,t) = P_0(n,t) + \sum_{j=1}^{N} \Omega^{-j/2} \sum_{m=1}^{3j} a_m^{(j)}(-\Omega^{1/2}\partial_n)^m P_0(n,t)$$
$$+ \, O(\Omega^{-(N+1)/2}). \tag{36}$$

The above follows from the definition of the eigenfunctions, Eq. (17), and using the derivative property of the convolution given by Eq. (34). Note that the coefficients in this equation are exactly the same as given in Eq. (18) and hence are determined by Eq. (22). One can verify two limiting cases: (i) as $\Sigma \to 0$ and $\Omega[X]$ being integer valued, then $P_0(n) = K(n - \Omega[X]) = \delta_{n,\Omega[X]}$ is just the Kronecker δ as required for deterministic initial conditions; (ii) as $\Omega \to \infty$ with y/Σ constant, the probability distribution P_0 reduces to the density π_0 given by Eq. (13) and hence it follows that in this limit the continuous and discrete series give the same results.

A. The discrete approximation performs well for linear birth-death processes

For the linear birth-death process in the previous section, in Fig. 1(b) we show that the discrete approximation given by Eq. (36) with Eq. (31) is in good agreement with the true distribution when truncated after the terms of order Ω^{-1} and shows no oscillations. This agreement is remarkable given the compact form of the solution given by Eq. (31) and (36). The approximation is almost indistinguishable from the exact result when the series is truncated after the Ω^{-3} terms using Eqs. (25) and (30) in Eq. (36). We hence conclude that the discrete series approximates better the underlying distribution of the master equation than the continuous approximation.

B. The discrete approximation fails for nonlinear birth-death processes

Next, we turn our attention to the analysis of nonlinear birth-death processes, i.e., a process whose propensities depend nonlinearly on the number of molecules. A particular feature

of such processes is that the LNA estimates for mean and variances are generally no longer exact but agree with those of the true distribution only in the limit of large system size [13].

As an example, we here consider a simple metabolic reaction confined in a small subcellular compartment of volume Ω with substrate input,

$$\varnothing \xrightarrow{h_0} S, \tag{37a}$$

$$S + E \underset{h_2}{\overset{h_1}{\rightleftharpoons}} C \xrightarrow{h_3} E. \tag{37b}$$
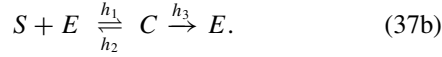
The reactions describe the input of substrate molecules S and their catalytic conversion by enzyme species E via the enzyme-substrate complex C. The SSE of the average concentrations correcting the macroscopic rate equations have been extensively studied [12]. Since our theory applies to a single species only, we here consider a reduced model in which reaction (37b) is modelled via an effective propensity: This gives $S_1 = +1$, $\gamma_1 = \Omega k_0$, and $S_2 = -1$, $\gamma_2 = \Omega k_1 \frac{n}{n+\Omega K}$. This simplification is valid when the enzyme-substrate association is in rapid equilibrium, which holds when $[E_T] \ll K$ and $h_3 \ll h_2$, where $[E_T]$ is the total enzyme concentration [19].

The parameters in the reduced model are related to those in the developed model by $k_0 = h_1$, $k_1 = h_3[E_T]$, and $K = h_2/h_1$. This reduced master equation is solved exactly by a negative binomial distribution [20].

The system size coefficients are obtained from Eq. (10) and are given by

$$\mathcal{D}_n^m = \delta_{m,0} k_0 + (-1)^n k_1 \frac{\partial^m}{\partial[X]^m} \frac{[X]}{K + [X]}, \tag{38}$$

and $\mathcal{D}_{n,s}^m = 0$ for $s > 0$. In Fig. 2(a) and 2(b), we consider two parameter sets corresponding to low and moderate numbers of substrate molecules, respectively. We observe that in contrast to the linear case, the discrete approximation of the nonlinear birth-death process tends to oscillate with increasing truncation order. This issue is addressed in the following section.

V. RENORMALIZATION OF NONLINEAR BIRTH-DEATH PROCESSES

Van Kampen's ansatz, Eq. (2), bears the particularly simple interpretation that for linear birth-death processes ϵ denotes the fluctuations about the average given by the solution of the rate
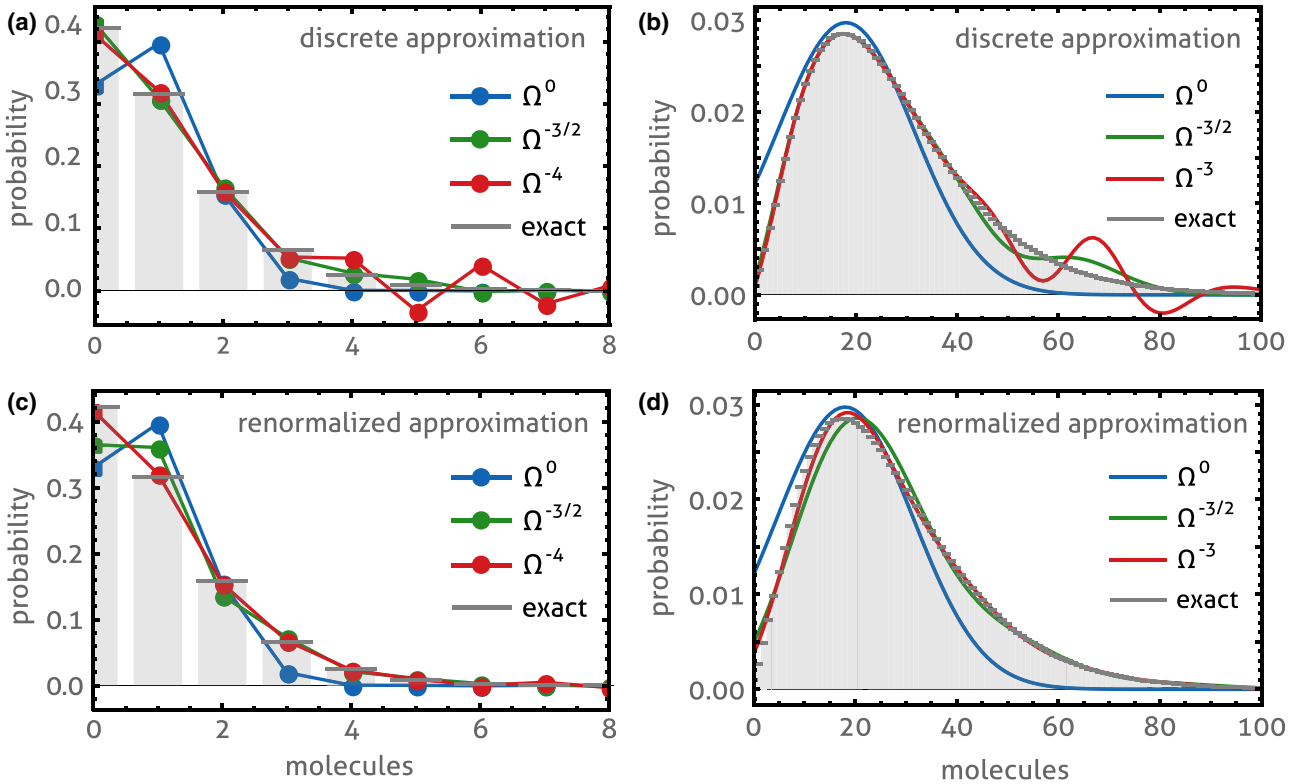

FIG. 2. (Color online) Nonlinear birth-death process. A metabolic reaction with Michaelis-Menten kinetics, scheme (37), is studied using the reduced model described in Sec. IV B. The exact stationary distribution is a negative binomial (shown in gray). (a) The discrete SSE approximation given by Eq. (36) with Eq. (25) and (30) is shown in the low-molecule-number regime ($k_0/k_1 = 0.25$, 1 molecule on average) when truncated after the Ω^0 (blue), $\Omega^{-3/2}$ (green), and Ω^{-4} terms (red dots). We observe that the expansion tends to oscillations and negative values of probability as the truncation order is increased. (b) Similar oscillations are observed for moderate molecule numbers ($k_0/k_1 = 0.9$, 27 molecules on average) for the discrete series truncated after the Ω^0 (blue), $\Omega^{-3/2}$ (green), and Ω^{-3} terms (red lines). In (c) and (d) we show the approximations corresponding to the same parameters used in (a) and (b), respectively, but obtained using the renormalization procedure given by Eq. (41) with Eq. (42) as described in the main text. The renormalized approximations avoid oscillations and are in excellent agreement with the true probability distributions (gray bars). We note that for the cases (b) and (d) the continuous and discrete approximations give essentially the same result. The remaining parameters are given by $\Omega = 10$ and $K = 0.1$.

equation $[X]$. As noted in the previous example, for nonlinear birth-death processes these estimates are only approximate. Their asymptotic series expansions will therefore require additional terms that compensate for the deviations of the LNA from the true concentration mean and variance. It would therefore be desirable to find an approximation for nonlinear processes that yields more accurate mean and variance than the LNA, for instance, by rewriting van Kampen's ansatz as

$$\frac{n}{\Omega} = \underbrace{[X] + \Omega^{-1/2}\langle\epsilon\rangle}_{\text{mean}} + \underbrace{\Omega^{-1/2}\bar{\epsilon}}_{\text{fluctuations}}. \tag{39}$$

Here $\bar{\epsilon} = \epsilon - \langle\epsilon\rangle$ denotes a centered variable that quantifies the fluctuations about the true average which is *a priori* unknown. These estimates, however, can be approximated using the SSE beforehand, and the asymptotic expansion of the distributions then can be performed about these new estimates. This idea is called renormalization and makes use of the fact that the terms correcting mean and variances can be summed exactly. As we show in the following the resummation allows us to better control the convergence by effectively reducing the number of terms in the summation while at the same time it retains the accuracy of the expansion.

The system size expansion of the moments, Eq. (24), yields the following estimates for mean and variance of the fluctuations:

$$\langle\epsilon\rangle = \sum_{j=0}^{N} \Omega^{-j/2} a_1^{(j)} + O(\Omega^{-(N+1)/2}), \tag{40a}$$

$$\bar{\sigma}^2 = \sigma^2 + \sum_{j=1}^{N} \Omega^{-j/2} \sigma_{(j)}^2 + O(\Omega^{-(N+1)/2}), \tag{40b}$$

respectively, where $\bar{\sigma}_{(j)}^2 = 2[a_2^{(j)} - B_{j,2}(\{\chi! a_1^{(\chi)}\}_{\chi=1}^{j-1})/j!]$ and $B_{j,n}$ are the partial Bell polynomials [21].

The renormalization procedure amounts to replacing y with $\bar{y} = (n - \Omega[X] - \Omega^{1/2}\langle\epsilon\rangle)$ and Σ^2 with $\bar{\Sigma}^2 = \Omega\bar{\sigma}^2$ in Eq. (35) and associating a new Gaussian $\bar{P}_0(n)$ with these estimates. The renormalized expansion is then given by

$$P(n,t) = \bar{P}_0(n,t) + \sum_{j=1}^{N} \Omega^{-j/2} \sum_{m=1}^{3j} \bar{a}_m^{(j)}(-\Omega^{1/2}\partial_n)^m \bar{P}_0(n,t)$$
$$+ O(\Omega^{-(N+1)/2}), \tag{41}$$

where the renormalized coefficients can be calculated from the bare ones using

$$\bar{a}_m^{(j)} = \sum_{k=0}^{j} \sum_{n=0}^{3k} a_n^{(k)} \kappa_{m-n}^{(j-k)} \tag{42}$$

and

$$\kappa_j^{(n)} = \frac{1}{n!} \sum_{m=0}^{\lfloor j/2\rfloor} (-1)^{(j+m)} \sum_{k=j-2m}^{n-m} \binom{n}{k}$$
$$\times B_{k,j-2m}\left(\{\chi! a_1^{(\chi)}\}\right) B_{n-k,m}\left(\left\{\frac{\chi!}{2}\bar{\sigma}_{(\chi)}^2\right\}\right), \tag{43}$$

where, again, $B_{k,n}(\{x_\chi\})$ denote the partial Bell polynomials [21]. The result is verified at the end of this section.

Note that the renormalized series has generally less nonzero coefficients since by construction $\bar{a}_1^{(j)} = \bar{a}_2^{(j)} = 0$. Note that for linear birth-death processes, mean and variance are exact to order Ω^0 (LNA), and hence for this case expansion (36) coincides with Eq. (41).

For example, truncating after the Ω^{-1} terms, from Eq. (40) it follows that $\langle\epsilon\rangle = \Omega^{-1/2}a_1^{(1)} + O(\Omega^{-3/2})$ and $\bar{\sigma}^2 = \sigma^2 + \Omega^{-1}[2a_2^{(2)} - (a_1^{(1)})^2] + O(\Omega^{-3/2})$. Using Eq. (42) the renormalized coefficients can be expressed in terms of the bare ones,

$$\bar{a}_1^{(1)} = 0, \quad \bar{a}_3^{(1)} = a_3^{(1)}, \tag{44a}$$

$$\bar{a}_2^{(2)} = 0, \quad \bar{a}_4^{(2)} = a_4^{(2)} - a_1^{(1)}a_3^{(1)}, \quad \bar{a}_6^{(2)} = a_6^{(2)}. \tag{44b}$$

This result can be used, for instance, to renormalize the stationary solution using the bare coefficients given in Sec. III E, Eqs. (27) and (28). The nonzero renormalized coefficients evaluate to

$$\bar{a}_3^{(1)} = -\frac{\sigma^4 \mathcal{D}_1^2}{6\mathcal{J}} + \frac{\sigma^2 \mathcal{D}_2^1}{6\mathcal{J}} + \frac{\mathcal{D}_3^0}{18\mathcal{J}}, \tag{45a}$$

$$\bar{a}_4^{(2)} = -\frac{\mathcal{D}_4^0}{96\mathcal{J}} - \frac{\sigma^2 \mathcal{D}_3^1}{24\mathcal{J}} - \frac{\sigma^4 \mathcal{D}_2^2}{16\mathcal{J}} - \frac{\sigma^6 \mathcal{D}_1^3}{24\mathcal{J}}$$
$$- \bar{a}_3^{(1)}\left(\frac{3\mathcal{D}_2^1}{8\mathcal{J}} + \frac{3\sigma^2 \mathcal{D}_1^2}{4\mathcal{J}}\right),$$

$$\bar{a}_6^{(2)} = \frac{1}{2}\big[\bar{a}_3^{(1)}\big]^2. \tag{45b}$$

Note that for linear birth-death processes $\mathcal{D}_{n,s}^m = 0$ for $s > 0$ and $m > 1$, and hence the above equations reduce to Eqs. (27) and (28).

A. The renormalized approximation performs well for nonlinear birth-death processes

For the metabolic reaction (37), mean and variance can be obtained to be $\langle\epsilon\rangle = \Omega^{-1/2}\varsigma + O(\Omega^{-2})$ and $\bar{\sigma}^2 = \sigma^2 + \Omega^{-1}\varsigma(\varsigma + 1) + O(\Omega^{-2})$, where $\varsigma = [X]/K$ is the reduced substrate concentration and $\sigma^2 = K\varsigma(\varsigma + 1)$. Substituting now Eq. (38) into Eqs. (45), we obtain the expansion coefficients

$$\bar{a}_3^{(1)} = \frac{\sigma^2}{6}(2\varsigma + 1), \tag{46a}$$

$$\bar{a}_4^{(2)} = \frac{\sigma^2}{24}[6\varsigma(\varsigma + 1) + 1], \quad \bar{a}_6^{(2)} = \frac{1}{2}\big(\bar{a}_3^{(1)}\big)^2, \tag{46b}$$

which determine the renormalized series expansion to order Ω^{-1}. Using Eq. (25), (38), and (42) we can give the next order terms to order $\Omega^{-3/2}$ analytically,

$$\bar{a}_3^{(3)} = \frac{\bar{a}_3^{(1)}}{K}, \quad \bar{a}_5^{(3)} = \frac{\bar{a}_3^{(1)}}{20}[12\varsigma(\varsigma + 1) + 1],$$

$$\bar{a}_7^{(3)} = \bar{a}_3^{(1)}\bar{a}_4^{(2)}, \quad \bar{a}_9^{(3)} = \frac{1}{6}(\bar{a}_3^{(1)})^3. \tag{46c}$$

In Figs. 2(c) and 2(d) we compare the renormalized approximation given by Eq. (41) with the respective bare approximations in Figs. 2(a) and 2(b). We observe that the

renormalization technique avoids oscillations and even the simple analytical approximation given by Eqs. (46) is in reasonable agreement with the exact result. We note that the asymptotic approximations shown in Figs. 2(c) and 2(d) are almost indistinguishable for higher truncation orders.

B. Proof of the renormalization formula

The renormalized coefficients can in principle be obtained by matching the expansions given by Eq. (36) and (41) via their characteristic functions. For convenience we consider the characteristic function of the series (18)

$$G(k) = G_0(k) \left[1 + \sum_{j=1}^{\infty} \Omega^{-j/2} \sum_{n=1}^{3j} a_n^{(j)} (ik)^n \right], \qquad (47)$$

with $G_0(k) = e^{-(k\sigma)^2/2}$ being the characteristic function solution of the LNA $\pi_0(\epsilon)$ and we have omitted the explicit time dependence to ease the notation. We are now looking for a different expansion with corrected estimates for the mean and variance.

$$\bar{G}(k) = \bar{G}_0(k) \left[1 + \sum_{j=1}^{\infty} \Omega^{-j/2} \sum_{n=1}^{3j} \bar{a}_n^{(j)} (ik)^n \right], \qquad (48)$$

Note that $\bar{G}_0(k) = e^{ik\langle\epsilon\rangle} e^{-(k\bar{\sigma})^2/2}$ is the characteristic function for a Gaussian random variable with mean $\langle\epsilon\rangle$ and variance $\bar{\sigma}^2$ given by Eqs. (40).

Equating now Eq. (47) and (48), we find

$$1 + \sum_{j=1}^{\infty} \Omega^{-j/2} \sum_{n=1}^{3j} \bar{a}_n^{(j)} (ik)^n$$
$$= \frac{G_0(k)}{\bar{G}_0(k)} \left[1 + \sum_{j=1}^{\infty} \Omega^{-j/2} \sum_{n=1}^{3j} a_n^{(j)} (ik)^n \right]. \qquad (49)$$

Expanding the prefactor in the above equation in powers of k and then in Ω, we have

$$\frac{G_0(k)}{\bar{G}_0(k)} = \sum_{j=0}^{\infty} (ik)^j \kappa_j = \sum_{n=0}^{\infty} \Omega^{-n/2} \sum_{j=0}^{2n} (ik)^j \kappa_j^{(n)}, \qquad (50)$$

from which Eq. (42) follows, which expresses the new coefficients $\bar{a}_n^{(j)}$ in terms of the bare ones $a_n^{(j)}$. It remains to derive an explicit expression for the $\kappa_j^{(n)}$. The expansion in powers of (ik) yields

$$\kappa_j = \sum_{m=0}^{\lfloor j/2 \rfloor} (-1)^{(j+m)} \frac{\langle\epsilon\rangle^{j-2m}}{(j-2m)!} \frac{\left(\frac{\bar{\sigma}^2 - \sigma^2}{2}\right)^m}{m!}. \qquad (51)$$

We now expand the first term in inverse powers of Ω using the partial Bell polynomials

$$\frac{1}{(j-2m)!} \left(\sum_{n=1}^{\infty} \Omega^{-n/2} a_1^{(n)} \right)^{j-2m}$$
$$= \sum_{n=1}^{\infty} \frac{\Omega^{-n/2}}{n!} \sum_{k=0}^{n} \delta_{j-2m,k} B_{n,k}\left(\left\{\chi! a_1^{(\chi)}\right\}\right), \qquad (52)$$
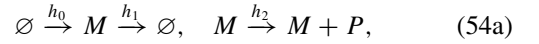
and similarly for the second term,

$$\frac{1}{m!} \left(\frac{1}{2} \sum_{n=1}^{\infty} \Omega^{-n/2} \bar{\sigma}_{(n)}^2 \right)^m$$
$$= \sum_{n=0}^{\infty} \frac{\Omega^{-n/2}}{n!} \sum_{k=0}^{n} \delta_{m,k} B_{n,k}\left(\left\{\frac{\chi!}{2} \bar{\sigma}_{(\chi)}^2\right\}\right). \qquad (53)$$

Using the above expansions in Eq. (51) and rearranging in powers of $\Omega^{-1/2}$, Eq. (43) for the coefficients $\kappa_j^{(n)}$ follows.

Finally, one associates with the centered variable $\bar{\epsilon} = \epsilon - \langle\epsilon\rangle$ a Gaussian $\bar{\pi}_0(\bar{\epsilon})$ with variance $\bar{\sigma}^2$. It then follows from inverting Eq. (48) that $\Pi(\bar{\epsilon}) = \bar{\pi}_0(\bar{\epsilon}) + \sum_{j=1}^{N} \Omega^{-j/2} \sum_{n=1}^{3j} \bar{a}_n^{(j)} \psi_n(\bar{\epsilon}) \bar{\pi}_0(\bar{\epsilon}) + O[\Omega^{-(N+1)/2}]$. Associating now the $\Omega^{-j/2}$ term of this equation with π_j in Eq. (33), the discrete series for $P(n,t)$ given by Eq. (41) follows.

VI. APPLICATION

The models studied so far have been useful to develop the method. It remains to be demonstrated, however, that it remains accurate in cases where analytical solution is not feasible, as, for instance, for out-of-steady-state and nondetailed balance systems. We here consider the synthesis of a protein P which is degraded through an enzyme

$$\varnothing \xrightarrow{h_0} M \xrightarrow{h_1} \varnothing, \quad M \xrightarrow{h_2} M + P, \qquad (54a)$$

$$P + E \underset{h_4}{\overset{h_3}{\rightleftharpoons}} C \xrightarrow{h_5} E, \qquad (54b)$$

where M denotes the transcript, E the enzyme, and C the complex species as studied in Ref. [22]. Since our theory applies only to a single species, we consider the limiting case in which the protein dynamics represents the slowest time scale of the system. It has been shown [23] that when species M is degraded much faster than the protein P, the protein synthesis (54a) reduces to the transition $S_1 = +z$, $\gamma_1 = \Omega k_0$ in which z is a random variable following the geometric distribution $\varphi(z) = \frac{1}{1+b}\left(\frac{b}{1+b}\right)^z$ with average b, which is called the burst approximation. Similarly to the metabolic reaction studied in Sec. IV, the enzymatic degradation process (54b) can be reduced to $S_2 = -1$, $\gamma_2 = \Omega k_1 \frac{n}{\Omega K + n}$ with a nonlinear dependence on the protein number n. The master equation describing the protein number is then given by

$$\frac{d}{dt} P(n) = \Omega \sum_{z=0}^{\infty} (E^{-z} - 1) k_0 \varphi(z) P(n)$$
$$+ \Omega (E^{+1} - 1) k_1 \frac{n}{\Omega K + n} P(n). \qquad (55)$$

The relation between the parameters in the reduced and the developed model are given by $k_0 = h_0 h_2/h_1$, $b = h_2/h_1$, $k_1 = h_5[E_T]$, $K = h_5/h_3$, where $[E_T]$ denotes the total enzyme concentration. This description involves countably many reactions: one for the degradation of the protein and one for each value of z. Therefore, the reactions cannot obey detailed balance in steady state. The system size coefficients now follow from Eq. (10) and are given by

$$\mathcal{D}_n^m = \delta_{m,0} k_0 \langle z^n \rangle_\varphi + (-1)^n k_1 \frac{\partial^m}{\partial [X]^m} \frac{[X]}{K + [X]}, \qquad (56)$$

and $\mathcal{D}_{n,s}^m = 0$ for $s > 0$, where $\langle z^n \rangle_\varphi = \sum_{z=0}^{\infty} z^n \varphi(z) = \frac{1}{1+b} \text{Li}_{-n}(\frac{b}{1+b})$ denotes the average over the geometric distribution in terms of the polylogarithm function [24]. The deterministic equation is given by

$$\frac{d[X]}{dt} = k_0 b - \frac{k_1[X]}{K + [X]}, \quad (57)$$

which follows from the expression for $\mathcal{D}_1^0$. Using the Jacobian $\mathcal{J} = \mathcal{D}_1^1$ and $\mathcal{D}_2^0$ in Eq. (14), we find that the LNA variance obeys

$$\frac{\partial \sigma^2}{\partial t} = -\frac{2k_1 K}{([X] + K)^2} \sigma^2 + k_0 b(1 + 2b) + \frac{k_1[X]}{K + [X]}. \quad (58)$$

The ODEs given by Eq. (57) and (58) are integrated numerically and the solution is used in Eq. (35) from which the leading-order approximation follows. Higher-order approximations are now be obtained by using Eq. (56) in (22) which govern the time evolution of the coefficients $a_m^{(j)}(t)$ and using the result in Eqs. (41) and (42). We assume deterministic initial conditions with zero proteins, meaning $a_m^{(j)}(0) = \delta_{m,0}\delta_{j,0}$. In Fig. 3(a) we compare the time evolution obtained by the leading-order approximation P_0 and Eq. (41) truncated after the Ω^{-3} term. The latter distributions are in excellent agreement with the distributions sampled using the stochastic simulation algorithm [25]. In particular, unlike the leading order approximation, these describe well mode, skewness, and tails of the distribution. We note that also the mean and variance of these distribution approximations are in excellent agreement as verified in inset of Fig. 3(a).

Despite the overall good agreement, in Fig. 3(b) we show that there are discrepancies at very short times where, again, the distribution approximations tend to oscillations. Motivated by this numerical observation, we speculate that this behavior of the expansion is due a temporal boundary layer as commonly observed in singular perturbation expansions [26]. Theoretically, the layer must be located at times of the same order as the expansion parameter, i.e., $t = (\Omega K)^{-1/2}$ min $\approx$ 13 s, coinciding with the simulation in Fig. 3(b). This suggests that our approach only describes the outer solution. Further analysis would be required to investigate also the inner solution which is beyond the scope of this article.

VII. DISCUSSION

We have here presented an approximate solution method for the probability distribution of the master equation. The solution is given in terms of an asymptotic series expansion that can be truncated systematically to any desired order in the inverse system size. For biochemical systems with large numbers of molecules, we have derived a continuous series approximation that extends van Kampen's LNA to higher orders in the SSE. In low-molecule-number conditions, we have found that this continuous approximation becomes inaccurate. Instead, in most practical situations, the prescribed discrete distribution approximations incorporating higher-order terms in the SSE better capture the underlying solution of the master equation. While the terms to order Ω^{-1} have been given explicitly, we found that for the examples studied here up to the Ω^{-3} or Ω^{-4} terms had to be taken into account to accurately characterize
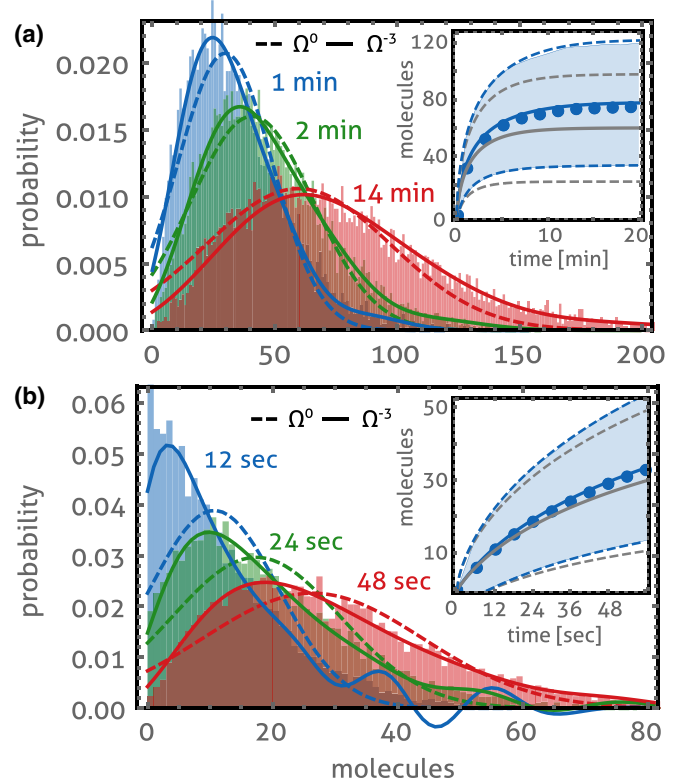

FIG. 3. (Color online) Predicting transient distributions of gene expression. The dynamics of protein synthesis with enzymatic degradation, scheme (54), is studied using the burst approximation (55). (a) We compare the time dependence of the renormalized discrete approximations to exact stochastic simulations at times 1, 2, and 14 min. The overall shape (mode, skewness, distribution tails) of the simulated distributions (bars) is in excellent agreement with the series approximation when truncated after the Ω^{-3} terms (solid lines) but not when only Ω^0 are taken into account (dashed lines). This agreement is also observed for the first two moments shown in the inset: While the Ω^{-3} approximation (blue solid line) agrees with the moment dynamics of the simulated distributions (dots) of the reduced model (55), the Ω^0 approximation underestimates the mean (gray solid line) and variance by 25%. The area within one standard deviation of the mean obtained from simulations is shown in blue, and the boundary obtained from the approximations are shown as dashed lines (Ω^0 gray, Ω^{-3} blue). (b) Despite the good agreement shown in (a) we found that at very short times (12 s, blue solid line) the series truncated after the Ω^{-3} terms tends to oscillations which quickly disappear for later times (24 s, green; 48 s, red solid line). See the main text for discussion. Parameters are $k_0\Omega = 15$ min^{-1}, $k_1\Omega = 100$ min^{-1}, $K\Omega = 20$, $\Omega = 100$, and $b = 5$. Histograms were obtained from 10 000 stochastic simulations.

these non-Gaussian distributions. We note, however, that the asymptotic expansion cannot generally guarantee the positivity of the probability law. These undulations are particularly pronounced in the short-time behavior of the expansion studied in Sec. VI, which our theory does not describe.

Previous means of solving the master equation have either been numerical in nature [27] or have focused on the inverse problem, i.e., reconstruction of the probability density from the moments. While a numerical solution for the master equation of a single species is rather straightforward, we expect our

procedure to become computationally advantageous when generalized to the multivariate case where numerical solution is usually prohibitive because of combinatorial explosion.

Methods based on moments typically require approximations such as moment closure [16] and also require the prior assumption of the first few moments containing all information on the probability distribution. Conversely, using the system size expansion, we have here obtained the probability distribution directly from the master equation *without the need to resort to moments*. This method enjoys the particular advantage over previous ones whereby the first few terms of this expansion can be written down explicitly as a function of the rate constants and for any number of reactions. For small models we have demonstrated that the procedure leads to particularly simple expressions for the non-Gaussian distributions. This development could prove particularly valuable for parameter estimation of biochemical reactions in living cells.

APPENDIX A: USEFUL PROPERTIES OF THE HERMITE POLYNOMIALS

We here briefly review some properties of the Hermite orthogonal polynomials. The polynomials can be defined in terms of the derivatives of a centered Gaussian π_0 with variance σ^2,

$$H_n\left(\frac{\epsilon}{\sigma}\right) = \pi_0^{-1}(\epsilon)(-\sigma\,\partial_\epsilon)^n \pi_0(\epsilon). \tag{A1}$$

An explicit formula is

$$H_n\left(\frac{\epsilon}{\sigma}\right) = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k}(-1)^k \left(\frac{\epsilon}{\sigma}\right)^{n-2k} (2k-1)!!. \tag{A2}$$

These functions are orthogonal $\frac{1}{n!}\int_{-\infty}^{\infty} d\epsilon\, H_m(\frac{\epsilon}{\sigma})H_n(\frac{\epsilon}{\sigma})\pi_0(\epsilon) = \delta_{nm}$, with respect to the Gaussian measure π_0. The derivative satisfies

$$(\sigma\,\partial_\epsilon)^m H_n\left(\frac{\epsilon}{\sigma}\right) = \frac{n!}{(n-m)!} H_{n-m}\left(\frac{\epsilon}{\sigma}\right). \tag{A3}$$

Since these polynomials are complete, every function $f(\epsilon)$ in $L_2(\mathbb{R}, \pi_0)$ (not necessarily positive) can be expanded as $f(\epsilon) = \sum_{n=0}^{\infty} b_n H_n(\frac{\epsilon}{\sigma})\pi_0(\epsilon)$, where the coefficients are given by $b_n = \frac{1}{n!}\int d\epsilon\, H_n(\frac{\epsilon}{\sigma})f(\epsilon)$. We note because $H_0(\frac{\epsilon}{\sigma}) = 1$ and π_0 is normalized, we must have $b_0 = 1$ if $\int d\epsilon\, f(\epsilon) = 1$.

APPENDIX B: EXPLICIT DERIVATION OF EQ. (23) AND THE PROPERTIES OF THE EXPANSION COEFFICIENTS

Changing variables $\epsilon = x\sigma$ and letting $\tilde{\mathcal{I}}_{mn}^{\alpha\beta} = \sigma^{\alpha-\beta+m-n}\mathcal{I}_{mn}^{\alpha\beta}$, the integral (20) can be written

$$\tilde{\mathcal{I}}_{mn}^{\alpha\beta} = \frac{1}{n!\alpha!\beta!}\int dx\, H_n(x)(-\partial_x)^\alpha x^\beta H_m(x)\pi_0(x), \tag{B1}$$

where $\pi_0(x)$ is a centered Gaussian with unit variance. Using partial integration, property (A3), and the relation

$$H_\alpha(x)H_\beta(x) = \alpha!\beta! \sum_{s=0}^{\min(\alpha,\beta)} \frac{H_{\alpha+\beta-2s}(x)}{s!(\alpha-s)!(\beta-s)!}, \tag{B2}$$

given in Ref. [28], one obtains

$$\tilde{\mathcal{I}}_{mn}^{\alpha\beta} = \frac{1}{\alpha!\beta!} \sum_{s=0}^{\min(n-\alpha,m)} \binom{m}{s} \frac{\int dx\, x^\beta H_{m+n-\alpha-2s}(x)\pi_0(x)}{(n-\alpha-s)!}. \tag{B3}$$

The remaining integral can now be evaluated as the moments of the unit Gaussian which yield

$$\int dx\, x^b H_a(x)\pi_0(x) = \frac{b!}{(b-a)!} \int dx\, x^{b-a}\pi_0(x)$$
$$= \frac{b!}{(b-a)!}(b-a-1)!! \tag{B4}$$

for even $(b-a) \geqslant 0$ and zero otherwise. Explicitly, the matrix elements are given by

$$\tilde{\mathcal{I}}_{mn}^{\alpha\beta} = \frac{1}{\alpha!} \sum_{s=0}^{\min(n-\alpha,m)} \binom{m}{s}$$
$$\times \frac{(\beta+\alpha+2s-(m+n)-1)!!}{[\beta+\alpha+2s-(m+n)]!(n-\alpha-s)!}, \tag{B5}$$

for even $(\alpha+\beta) - (m+n)$ and zero otherwise. Note that the above quantity is strictly positive. Note also that the argument of the double factorial is taken to be positive and hence the summation is nonzero only if $\alpha + \beta + 2\min(n-\alpha,m) \geqslant m+n$ and hence for even $\beta = 2k$ we have $n = m + \alpha \pm 2l$, while for odd $\beta = (2k+1)$ we have $n = m + \alpha \pm (2l+1)$, with $l = 0,\ldots,k$.

The integral formula can be used to verify two important properties of the solution of Eq. (22) given deterministic initial conditions: (i) We have $N_j = 3j$ and hence Eq. (22) indeed yields a finite number of equations. (ii) The coefficients $a_n^{(j)}$ for which $(n+j)$ is odd vanish at all times.

To verify property (i), let N_j be the index of the highest eigenfunction required to order $\Omega^{-j/2}$. Using Eq. (22) one can show that $a_{N_j}^{(j)} \sim a_{N_{j-1}}^{(j-1)}\mathcal{I}_{N_{j-1},N_j}^{p,3-p}$ for $p \in \{1,2,3\}$. By virtue of the properties given after Eq. (23), we find $N_j = N_{j-1} + 3$. Since for deterministic initial conditions we have $N_0 = 0$, it follows that $N_j = 3j$.

Finally, we verify property (ii). To the summation in Eq. (22) there contribute only terms for which $\mathcal{I}_{mn}^{p,k-p-2(s-1)}$ is nonzero. Hence, by the condition given after Eq. (23), $k - (m+n)$ is an even number. Considering the equation for $a_n^{(j)}$ for which $n + j$ is even, it follows that in the summation on the right-hand side of Eq. (22) there appear only coefficients for which $m + (j-k)$ is even. Conversely, for $n + j$ being odd, $m + (j-k)$ must be odd. Hence the pairs of equations for $a_n^{(j)}$ for which $(j+n)$ is even or odd are mutually uncoupled. For deterministic initial conditions, only terms with $j+n$ even differ from zero initially from which the result follows.

APPENDIX C: SOLUTION TO THE PROBLEM OF MOMENTS USING THE SYSTEM SIZE EXPANSION

Having obtained the moment expansion in terms of the coefficients $a_n^{(j)}$, it would be desirable to invert this relation and the coefficients in terms of the expansion of the moments. This can be derived using the completeness of the Hermite polynomials and writing the probability density as $\Pi(\epsilon) = \sum_{n=0}^{\infty} b_n H_n(\frac{\epsilon}{\sigma})\pi_0(\epsilon)$, where the $b_n = \frac{1}{n!}\int d\epsilon H_n(\frac{\epsilon}{\sigma})\Pi(\epsilon)$ can be expressed in terms of the moments using Eq. (A2), as follows:

$$b_n = \frac{1}{n!}\sum_{k=0}^{\lfloor n/2 \rfloor}\binom{n}{2k}(-1)^k\frac{\langle \epsilon^{n-2k} \rangle}{\sigma^{n-2k}}(2k-1)!!. \qquad (C1)$$

Assuming now that the moments can be expanded in a series in powers of Ω, i.e.,

$$\langle \epsilon^\beta \rangle = \sum_{j=0}^{N}\Omega^{-j/2}[\epsilon^\beta]_j + O(\Omega^{-(N+1)/2}), \qquad (C2)$$

the b_n can be matched to the coefficients a_n in Eq. (18) using $\sigma^n b_n = \sum_{j=0}^{N}\Omega^{-j/2}a_n^{(j)} + O[\Omega^{-(N+1)/2}]$, from which

one obtains

$$a_n^{(j)} = \frac{1}{n!}\sum_{k=0}^{\lfloor n/2 \rfloor}\binom{n}{2k}(-\sigma^2)^k(2k-1)!![\epsilon^{n-2k}]_j, \qquad (C3)$$

with $[\epsilon^0]_j = \delta_{j,0}$. The above formula relates the expansion of the moments to the expansion of distribution functions. It is now evident that the system size expansion of the distribution can be constructed from the system size expansion for a finite set of moments.

Specifically, to order $\Omega^{-1/2}$ the nonzero coefficients evaluate to

$$a_1^{(1)} = [\epsilon]_1, \quad a_3^{(1)} = \frac{1}{3!}\big([\epsilon^3]_1^3 - 3\sigma^2[\epsilon]_1\big), \qquad (C4)$$

while the coefficients to order Ω^{-1} are given by

$$a_2^{(2)} = \frac{1}{2}[\epsilon^2]_2, \quad a_4^{(2)} = \frac{1}{4!}([\epsilon^4]_2 - 6\sigma^2[\epsilon^2]_2),$$

$$a_6^{(2)} = \frac{1}{6!}(45\sigma^4[\epsilon^2]_2 - 15\sigma^2[\epsilon^4]_2 + [\epsilon^6]_2). \qquad (C5)$$

A different series is obtained using the Edgeworth expansion which, instead of using the system size expansion of the moments, Eq. (C2), proceeds by scaling the cumulants by a size parameter.

[1] T. Jahnke and W. Huisinga, J. Math. Biol. **54**, 1 (2007).
[2] H. Haken, Phys. Lett. A **46**, 443 (1974); N. G. Van Kampen, *ibid.* **59**, 333 (1976).
[3] R. M. Mazo, J. Chem. Phys. **62**, 4244 (1975); J. Peccoud and B. Ycart, Theor. Popul. Biol. **48**, 222 (1995); P. Bokes, J. R. King, A. T. Wood, and M. Loose, J. Math. Biol. **64**, 829 (2012).
[4] R. Görtz and D. Walls, Z. Phys. B Con. Mat. **25**, 423 (1976); G. Haag and P. Hänggi, *ibid.* **34**, 411 (1979).
[5] D. Schnoerr, G. Sanguinetti, and R. Grima, J. Chem. Phys. **141**, 024103 (2014).
[6] N. G. Van Kampen, Adv. Chem. Phys. **34**, 245 (1976); *Stochastic Processes in Physics and Chemistry*, 3rd ed. (Elsevier, Amsterdam, 1997).
[7] J. Elf and M. Ehrenberg, Genome Res. **13**, 2475 (2003).
[8] A. Melbinger, L. Reese, and E. Frey, Phys. Rev. Lett. **108**, 258104 (2012); J. Szavits-Nossan, K. Eden, R. J. Morris, C. E. MacPhee, M. R. Evans, and R. J. Allen, *ibid.* **113**, 098101 (2014).
[9] G. Rozhnova and A. Nunes, Phys. Rev. E **79**, 041922 (2009).
[10] M. Aoki, *Modeling Aggregate Behavior and Fluctuations in Economics* (Cambridge University Press, Cambridge, 2001).
[11] T. Heskes, J. Phys. A: Math. Gen. **27**, 5145 (1994).
[12] R. Grima, Phys. Rev. Lett. **102**, 218103 (2009); P. Thomas, A. V. Straube, and R. Grima, J. Chem. Phys. **133**, 195101 (2010).
[13] R. Grima, P. Thomas, and A. V. Straube, J. Chem. Phys. **135**, 084103 (2011).
[14] C. Cianci, F. Di Patti, D. Fanelli, and L. Barletti, Eur. Phys. J. Special Topics **212**, 5 (2012); P. Thomas, C. Fleck, R. Grima, and N. Popovic, J. Phys. A: Math. Theor. **47**, 455007 (2014).
[15] S. Engblom, Appl. Math. Comput. **180**, 498 (2006); R. Grima, J. Chem. Phys. **136**, 154105 (2012); A. Ale, P. Kirk, and M. P. Stumpf, *ibid.* **138**, 174101 (2013).
[16] V. Sotiropoulos and Y. N. Kaznessis, Chem. Eng. Sci. **66**, 268 (2011); P. Smadbeck and Y. N. Kaznessis, Proc. Natl. Acad. Sci. USA **110**, 14261 (2013); A. Andreychenko, L. Mikeev, and V. Wolf, ACM Trans. Model. Comput. Simul. **25**, 12 (2015).
[17] C. Cianci, F. Di Patti, and D. Fanelli, Europhys. Lett. **96**, 50011 (2011).
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (John Wiley & Sons, Hoboken, NJ, 2012).
[19] P. Thomas, A. V. Straube, and R. Grima, J. Chem. Phys. **135**, 181103 (2011); K. R. Sanft, D. T. Gillespie, and L. R. Petzold, IET Syst. Biol. **5**, 58 (2011).
[20] J. Paulsson, O. G. Berg, and M. Ehrenberg, Proc. Natl. Acad. Sci. U.S.A. **97**, 7148 (2000).
[21] L. Comtet, *Advanced Combinatorics*, 1st ed. (Springer, Netherlands, 1974); The partial Bell polynomials are defined as $B_{n,k}(\{x_\chi\}) = \sum' \frac{n!}{j_1!..j_{n-k+1}!}(\frac{x_1}{1!})^{j_1}..(\frac{x_{n-k+1}}{(n-k+1)!})^{j_{n-k+1}}$, where the summation $\sum'$ is such that $j_1 + \ldots + j_{n-k+1} = k$ and $j_1 + 2j_2 + \ldots + (n-k+1)j_{n-k+1} = n$ and $\{x_\chi\}$ denotes the sequence $(x_1, x_2, x_3, \ldots)$. These are available in Mathematica via the function BellY .
[22] P. Thomas, H. Matuschek, and R. Grima, in *IEEE International Conference on Bioinformatics Biomed (BIBM), 2012* (IEEE, Piscataway, NJ, 2012), pp. 1–5.
[23] V. Shahrezaei and P. S. Swain, Proc. Natl. Acad. Sci. USA **105**, 17256 (2008).
[24] The polylogarithm is defined by $\mathrm{Li}_s(x) = \sum_{k=1}^{\infty}\frac{x^k}{k^s}$ and implemeted in Mathematica by PolyLog $[x, s]$.

[25] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977); exact realizations of the process described by Eq. (55) are obtained using the Gillespie's algorithm with the increments of the synthesis reaction being replaced by independently and identically geometrically distributed integers of mean b.

[26] F. Verhulst, *Methods and Applications of Singular Perturbations* (Springer, New York, 2006).

[27] B. Munsky and M. Khammash, J. Chem. Phys. **124**, 044104 (2006).

[28] N. N. Lebedev, *Special Functions and Their Applications* (Dover, New York, 1972).