Approximation of stochastic birth-death-mutation processes with the Kramers-Moyal expansion

Pranav Murugan*
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge MA
(Dated: May 19, 2023)

Many complex biological phenomena can be modeled as chemical reaction networks, including common evolutionary processes. Birth-death-mutation processes are a key component of many of these evolutionary models, and can be fully described a (highly complex) master equation. In this work, we use the Kramers-Moyal expansion to convert the master equation into a much simpler partial differential equation, which can then be integrated numerically. We find that keeping additional terms in the Kramers-Moyal expansion has negligible benefit, and that such approximations tend to overestimate extinction rates at longer times.

I. INTRODUCTION

Biological systems are governed by a complex and diverse array of physical processes. These systems are often far from equilibrium and can be challenging to fully specify and understand [1]. This is indeed true for evolutionary processes, which couple together the death and replication of multiple species which can also interact and experience a time-varying environment. The physics of such a phenomenon is difficult to exactly define, but many approximate models have been designed that capture some of the important dynamics of these biological systems.

Some of these models operate in the mean-field limit, describing the deterministic dynamics of macroscopic observables such as population means [2, 3]. However, when we are not operating in the large population limit as is often the case when investigating rare mutants or cell types, the apparent stochasticity of the processes of replication and death must be included in our model. One popular way is to describe the birth-death-mutation processes analogously to chemical reactions, where each event occurs with a given (but potentially time-varying) propensity, where the rate of transitioning between states are given by the master equation [4, 5]. Master equations describing population dynamics have been applied widely in the context of evolutionary problems in biology [6–8]. This model of biological systems as networks of reacting particles is broad and powerful, but is challenging to interrogate analytically except for the simplest cases while computational efforts scale poorly with the size of the system [9, 10]. However, techniques from statistical physics have been used to develop new approaches to get estimates of system observables in a much more tractable way.

In this paper, we investigate two methods by which we can obtain approximate solutions to various properties of systems described by stochastic master equations. First is the Kramers-Moyal expansion of the discrete master equation, which approximates a large system of coupled ordinary differential equation as a single partial differential equation of arbitrary order [4]. We show that including additional terms in the Kramers-Moyal expansion does not contribute significantly to accuracy, and we highlight regimes where this approximation performs better than direct sampling [11].

II. BIRTH-DEATH PROCESSES

The simplest reaction network we will study is the stochastic birth-death process, which consists of the reactions

$$X \xrightarrow[\text{birth}]{\rho} X + X, \quad X \xrightarrow[\text{death}]{\lambda} \varnothing, \tag{1}$$

for reaction rates ρ and λ. For mass-action kinetics, as is often assumed, the propensity of a reaction is the reaction rate weighted by the degeneracy of the reaction – if there are x copies of X, for example, the rate at which a birth event occurs is ρx. The master equation for these reactions describe the probability p of having n copies of X at time t, and are given by

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} =& -[\rho+\lambda]\,xp(x,t) \\ &+ \rho(x-1)p(x-1,t) \\ &+ \lambda(x+1)p(x+1,t).\end{aligned} \tag{2}$$

We can sample trajectories of x over time directly from (1) and (2) using variants of Gillespie's stochastic simulation algorithm [11]. However, our goal is to find a more computationally feasible method of approximating the solution to the master equations, so we perform the Kramers-Moyal expansion. Heuristically, we assume $x \gg 1$ and perform a Taylor series expansion of (2)

* pmurugan@mit.edu

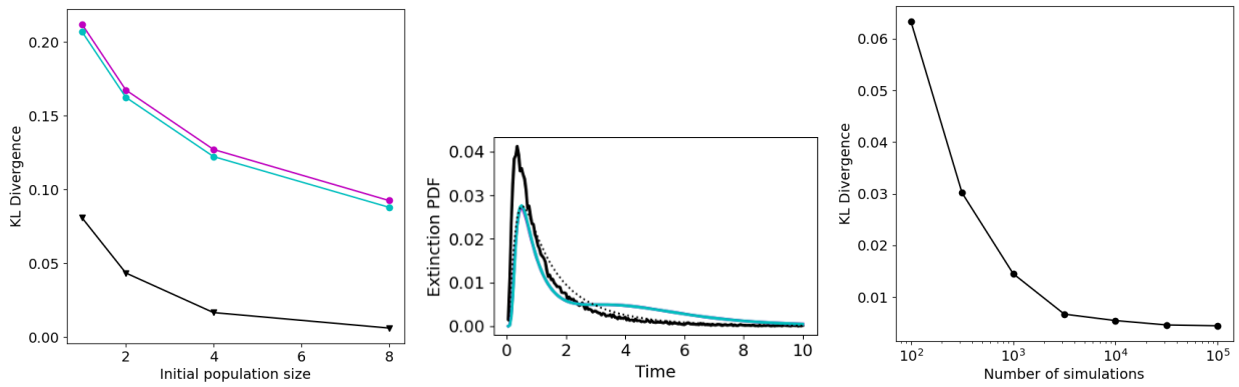

FIG. 1. **Left.** The KL divergence between the extinction PDF calculated between the ground truth reference and integration of (3) with 2 (magenta) and 3 (cyan) terms, as well as stochastic simulation of (2) as a function of x at $t = 0$. **Middle.** Extinction probability density function with initial conditions $x = 4, t = 0$ for (3) with 2 (magenta) and 3 (cyan) terms, stochastic simulation of (2) (black solid), and the analytical result (black dotted). **Right.** KL divergence of the extinction PDF estimated from stochastic simulation of the master equation as a function of the number of simulations.

around x to get

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} =& -(\rho-\lambda)\frac{\partial[xp(x,t)]}{\partial x} \\ &+\left(\frac{\rho+\lambda}{2}\right)\frac{\partial^2[xp(x,t)]}{\partial x^2} \\ &-\left(\frac{\rho-\lambda}{6}\right)\frac{\partial^3[xp(x,t)]}{\partial x^3}+\dots \end{aligned} \tag{3}$$

While this approach is not fully rigorous, it has been shown that more formal techniques such as the system size expansion yield the same results to low orders [4]. Termination of the right hand side of (3) at the second-order derivative yields the familiar form of the Fokker-Planck equation [12].

To compare the accuracy of these different methods, we compute the extinction probability density as a function of time (i.e. $\frac{dp(0,t)}{dt} \equiv q(t)$) and compare to the analytical result [5]. Our choice of the extinction probability density function (PDF) is motivated by the fact that it is easy to interpret and is sensitive to accuracy in the non-deterministic (small population) regime. The error with respect to the analytical ground truth can easily be computed using the KL divergence

$$D_{KL}(p^A||p^B) = \int dt\, p^A(t)\log\left(\frac{p^A(t)}{p^B(t)}\right) \tag{4}$$

where p^A is the test probability distribution and p^B is the reference probability distribution. We note that the exact solution of the extinction PDF as a function of time for a linear birth-death process is known, and is used in this section as the reference distribution [5].

We integrate (3) with an absorbing boundary condition at $x = -0.1$ (the discretization size of the integration grid) until $t = 10$, for $\rho = 5$ and $\lambda = 4.5$. We compare the results of the various methods for different initial conditions $p(x, 0)$, as seen in Figure 1. We see that the accuracy of direct stochastic simulation (in this case, with 50000 simulations) is more accurate to the ground truth than the Kramers-Moyal expansion. Interestingly, adding a third term only slightly improves the accuracy of the extinction time PDF, which suggests that higher-order terms may not contribute much to the accuracy of the expansion at small x (where the assumptions of the expansion do not hold). Thus, it also makes sense that the accuracy of the Kramers-Moyal methods increase as the initial population size increases, although as we can see in the middle figure in Figure 1, the Kramers-Moyal-derived methods appear to overestimate the probability density at later times. Finally, at least in this simple 1-d birth-death model, we see that the Gillespie stochastic simulation of the master equations converges very quickly to the ground truth PDF, in less than 10000 simulations.

III. LOGISTIC GROWTH

The simple birth-death model in the previous section can be made more realistic by adding a logistic growth term, which takes the form of a quadratic death rate:

$$X \xrightarrow[\text{logistic}]{\Lambda x} \varnothing, \tag{5}$$

where the reaction propensity is Λx^2. Such a term models resource limitation present in real biological systems, which determines the carrying capacity of the population. From the mean-field formulation of these dynamics, we can approximate the steady-state population of the system as $\tilde{x} = (\rho-\lambda)/\Lambda$ for $\rho > \lambda$ [13]. The master equation

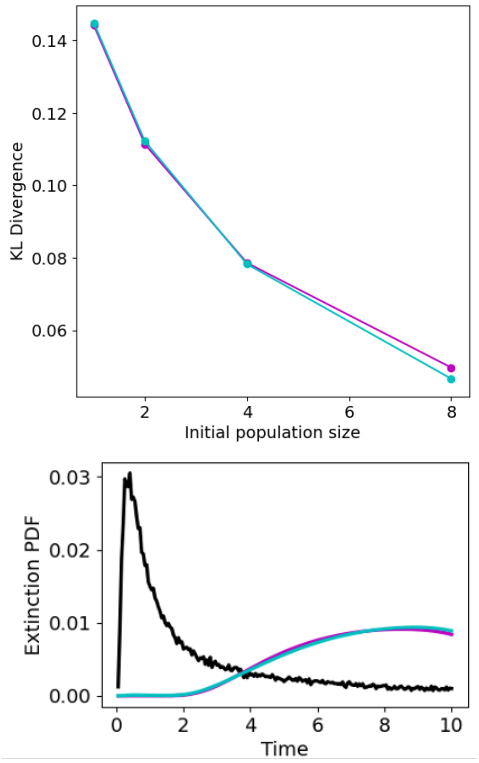

FIG. 2. **Top.** The KL divergence between the extinction PDF calculated from integration of (3) with 2 (magenta) and 3 (cyan) terms with respect to the reference obtained from stochastic simulation of (6) as a function of x at $t = 0$. **Bottom.** Extinction probability density function with initial conditions $x = 4, t = 0$ for (7) with 2 (magenta) and 3 (cyan) terms, and stochastic simulation of (6) (black solid).

including logistic growth is

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} =& -\left[\rho+\lambda+\Lambda x\right] x p(x,t) \\ &+ \rho(x-1)p(x-1,t) \\ &+ \left[\lambda+\Lambda(x+1)\right](x+1)p(x+1,t)\end{aligned} \tag{6}$$

which has corresponding Kramers-Moyal expansion

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} =& -\frac{\partial}{\partial x}\left[(\rho-\lambda)x-\Lambda x^2\right]p(x,t) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[(\rho-\lambda)x-\Lambda x^2\right]p(x,t) \\ &- \frac{1}{6}\frac{\partial^3}{\partial x^3}\left[(\rho-\lambda)x-\Lambda x^2\right]p(x,t) \\ &+ \dots\end{aligned} \tag{7}$$

As in the previous section, we compute the extinction PDFs for both the degree-two and degree-three approximations from the Kramers-Moyal expansion (7). We see in Figure 2 that we observe similar trends as in the basic birth-death case; KL divergence decreases with starting population, and keeping higher order terms does not appreciably change the predicted PDF. Interestingly, the bottom plot of Figure 2 indicates that the Kramers-Moyal approximation overestimates the extinction PDF much worse with the logistic growth term than without. This characteristic overestimation may be due to the finite upper boundary necessary to integrate a partial differential equation. The proper boundary condition is a Neumann condition at infinity, but in the solver is incorporated at a finite value ($x = 100$). This reflective boundary condition may artificially inflate the extinction probability and is simply an artifact of limited computational power.

IV. BIRTH-DEATH-MUTATION PROCESSES

We can extend the Kramers-Moyal formalism to more complex systems. We now consider the case where there are multiple species X_i, each of which can replicate and die with species-specific rates, but which also have the ability to mutate – in other words, we add the following class of reactions:

$$X_i \xrightarrow[\text{mutation}]{\mu} X_j, \tag{8}$$

where $i \neq j$ index the different species. We can encode the state of the system in the vector $\mathbf{x}$, where each element x_i indicates the number of species of type i that exist. The master equation for this process then takes the form

$$\begin{aligned}\frac{\partial p(\mathbf{x},t)}{\partial t} =& \sum_i \{-\left[\rho_i+\lambda_i+\mu\right]x_i p(\mathbf{x},t) \\ &+ \rho_i(x_i-1)p(\mathbf{x}-\{1\}^i,t) \\ &+ \lambda_i(x_i+1)p(\mathbf{x}+\{1\}^i,t) \\ &+ \sum_{j\neq i}\mu(x_i+1)p(\mathbf{x}+\{1\}^i-\{1\}^j,t)\end{aligned} \tag{9}$$

where $\{1\}^i$ is the vector of 0s with a 1 at index i. Of course, in general, μ can be a function of i or j; for simplicity we consider the case of constant μ. The Fokker-Planck approximation to these dynamics (from the Kramers-Moyal expansion of degree two) is then:

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = \sum_i \left\{ -(\rho_i - \lambda_i)\partial_i[x_i p(\mathbf{x},t)] + \frac{\rho_i + \lambda_i}{2}\partial_i^2 \left[x_i p(\mathbf{x},t)\right] + \mu \left(\partial_i \left[x_i p(\mathbf{x},t)\right] - \sum_{j\neq i} x_i \partial_j \left[p(\mathbf{x},t)\right] \right) \right.$$
$$\left. + \frac{\mu}{2}\left(\partial_i^2 \left[x_i p(\mathbf{x},t)\right] + \sum_{j\neq i} x_i \partial_j^2 \left[p(\mathbf{x},t)\right] - 2\partial_i\partial_j \left[x_i p(\mathbf{x},t)\right] \right) + \dots \right\}. \tag{10}$$

This equation, in principle, can be integrated over time to obtain the joint probability of the species population. However, such integration takes some care to numerically perform due to computational limitations; such optimizations were beyond the scope of this project, but can be done in principle as in [12]. We include this general computation of the Kramers-Moyal expansion of the birth-death-mutation process to demonstrate its similarity to the methods in sections II and III.

V. DISCUSSION

Biological reaction networks are a powerful model that can be used to describe many physical systems. However, in many cases these models can be difficult to simulate and impossible to analytically solve. In this paper we expand on the Kramers-Moyal method to convert the master equation describing the reaction network, which is in principle a system of coupled ordinary differential equations, into a much more tractable partial differential equation. We find that truncating the expansion after three terms has negligible benefit over truncation after two, and we find in general that the Kramers-Moyal approximation works best in the regime of large population sizes and tends to overestimate probabilities of extinction. However, in the regimes where the approximation works well, we can look to combine the benefits of solving the Fokker-Planck equation (e.g. well-defined support, easy physical interpretation) with the more accurate results of stochastic simulation.

[1] X. Fang, K. Kruse, T. Lu, and J. Wang, Nonequilibrium physics in biology, Reviews of Modern Physics **91**, 10.1103/revmodphys.91.045004 (2019).

[2] G. Sella and A. E. Hirsh, The application of statistical physics to evolutionary biology, Proceedings of the National Academy of Sciences **102**, 9541 (2005).

[3] D. B. Saakian and C.-K. Hu, Exact solution of the eigen model with general fitness functions and degradation rates, Proceedings of the National Academy of Sciences **103**, 4935 (2006).

[4] C. W. Gardiner, en*Stochastic methods*, Springer Series in Synergetics (Springer, Berlin, Germany, 2010).

[5] L. J. S. Allen, *An introduction to stochastic processes with applications to biology*, 2nd ed. (Whittles Publishing, Caithness, UK, 2010).

[6] R. S. Ganti and A. K. Chakraborty, Mechanisms underlying vaccination protocols that may optimally elicit broadly neutralizing antibodies against highly mutable pathogens, Physical Review E **103**, 10.1103/physreve.103.052408 (2021).

[7] D. A. Birch and W. R. Young, A master equation for a spatial population model with pair interactions, Theoretical Population Biology **70**, 26 (2006).

[8] S. Leibler and E. Kussell, Individual histories and selection in heterogeneous populations, Proceedings of the National Academy of Sciences **107**, 13183 (2010).

[9] B. J. Daigle, M. K. Roh, D. T. Gillespie, and L. R. Petzold, Automated estimation of rare event probabilities in biochemical systems, The Journal of Chemical Physics **134**, 044110 (2011).

[10] J. J. Vastola, Solving the chemical master equation for monomolecular reaction systems and beyond: a doi-peliti path integral view, Journal of Mathematical Biology **83**, 10.1007/s00285-021-01670-7 (2021).

[11] D. T. Gillespie, Stochastic simulation of chemical kinetics, Annual Review of Physical Chemistry **58**, 35 (2007).

[12] P. Sjöberg, P. Lötstedt, and J. Elf, Fokker–planck approximation of the master equation in molecular biology, Computing and Visualization in Science **12**, 37 (2007).

[13] R. E. Baker and M. J. Simpson, Correcting mean-field approximations for birth-death-movement processes, Physical Review E **82**, 10.1103/physreve.82.041905 (2010).