A derivation of the master equation from path entropy maximization

Julian Lee[1,a)] and Steve Pressé[2,b)]
[1]*Department of Bioinformatics and Life Science, Soongsil University, Seoul, South Korea*
[2]*Physics Department, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, Indiana 46202, USA*

(Received 7 June 2012; accepted 25 July 2012; published online 15 August 2012)

The master equation and, more generally, Markov processes are routinely used as models for stochastic processes. They are often justified on the basis of randomization and coarse-graining assumptions. Here instead, we derive *n*th-order Markov processes and the master equation as unique solutions to an inverse problem. We find that when constraints are not enough to uniquely determine the stochastic model, an *n*th-order Markov process emerges as the unique maximum entropy solution to this otherwise underdetermined problem. This gives a rigorous alternative for justifying such models while providing a systematic recipe for generalizing widely accepted stochastic models usually assumed to follow from the first principles. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4743955]

I. INTRODUCTION

Markov chains[1,2] are often the starting point for modeling condensed phase stochastic dynamics in biophysics[3–8] and beyond.[9] Markov chains are approximations of continuous system dynamics. They are primarily justified on the basis of coarse-graining approximations.[10] Coarse-graining reduces classical phase space—with phase points dynamics governed by Liouville's equations—to a discrete set of states—with stochastic hopping between states determined by stationary transition probabilities. Such coarse-graining methods have recently been used to show how Markov models can describe the continuous dynamics of biomolecules evolving in complex potential landscapes.[11–13]

A very different approach to stochastic dynamics is due to Filyukov and Karpov[14] and later Jaynes.[15] Using this approach, stochastic dynamical models can be inferred as unique solutions to an inverse problem. A model is defined by the probability for each stochastic path. Normally, the number of stochastic paths greatly outnumbers the constraints imposed from data. To find a unique solution to this underdetermined problem we ask: which model not only satisfies the limited experimental constraints but also maximizes the entropy for the path probabilities? This is exactly equivalent to finding a model for the path probabilities which satisfies the experimental constraints while satisfying these logical consistency axioms due to Shore and Johnson:[16] (1) when A and B are independent data then the model for $P(A \text{ and } B)$ must reduce to $P(A)P(B)$ and the model for $P(A \text{ or } B)$ must reduce to $P(A) + P(B)$; and (2) furthermore, any prediction made from the model must be independent of the coordinate system used in the calculation.

This method of finding a stochastic model is mathematically similar to the maximum entropy principle for determining equilibrium probability distributions.[17–20] In earlier work, Ge *et al.*—which extended the work of Stock *et al.*[21] and Ghosh *et al.*[22]—showed that the first order Markov chain emerges as a natural consequence of path entropy maximization. Here we generalize this work in many important ways. (1) We do not limit ourselves to first order Markov processes; (2) we consider the conditions for which the master equation emerges as a solution to the procedure of path entropy maximization; (3) we consider how different types of constraints affect the emergent model; and (4) we consider very general (nonlinear) constraints.

To the best of our knowledge, this is the first time the master equation and, more generally, *n*th-order Markov processes are rigorously shown to follow from maximum entropy principles. This provides an alternative justification for the master equation—the basic tool of stochastic physics and biology—which is distinct from standard chemical or mechanistic justifications provided by van Kampen,[1] Zwanzig,[23] Gillespie,[24] and others. The master equation assumes from the onset a dynamics described by stationary transition probabilities and time-varying state occupation probabilities. Here we only assume data of a specific type are available and the basic logical consistency axioms required to justify maximum entropy as an inference tool.[16] Posing the master equation as the solution of an inverse problem is significant because possible generalizations to the master equation are now derivable within this formalism. These generalizations can then be justified on the firm axiomatic basis of provided by Shore and Johnson.

II. MARKOV MODEL OF *n*th ORDER: DEFINITIONS AND NOTATIONS

In this section, we briefly introduce the mathematical notation necessary for the remainder of the paper. Consider a stochastic process in discrete time. Let the index i_t denote the state of the system at time t along the path C from time 0 to T where $C = \{i_0, i_1, i_2, \ldots, i_T\}$. The probability distribution of

a)Electronic mail: jul@ssu.ac.kr.
b)Electronic mail: stevenpresse@gmail.com.

path C is

$$P(C) = p(i_0, i_1, \ldots, i_T). \tag{1}$$

An n-point joint probability is defined as follows

$$\begin{aligned} &p(a_1, \ldots, a_m; t) \\ &\equiv \sum_{i_0, i_1, \ldots, i_{t-m}, j_1, j_2, \ldots, j_{T-t}} p(i_0, i_1, \ldots, i_{t-m}, \\ &\quad a_1, \ldots, a_m, j_1, j_2, \ldots, j_{T-t}). \end{aligned} \tag{2}$$

The explicit time index is required, as the result depends on which indices are summed over. Conditional—also called transition—probabilities are obtained by invoking Bayes' theorem:

$$p(i_0, \ldots, i_{t-1} \to i_t) \equiv \frac{p(i_0, \ldots, i_t)}{p(i_0, \ldots, i_{t-1})}. \tag{3}$$

We call $p(i_0, \ldots, i_{t-1} \to i_t)$ a transition probability. When the transition probability depends only on the previous n-time steps

$$\begin{aligned} p(i_0, \ldots, i_{t-1} \to i_t) &= p(i_{t-n}, i_{t-n+1} \ldots, i_{t-1} \to i_t; t) \\ &\equiv \frac{p(i_{t-n}, \ldots, i_t; t)}{p(i_{t-n}, \ldots, i_{t-1}; t-1)}, \end{aligned} \tag{4}$$

the process is called an *nth-order Markov process*. When the transition probability is time-independent, it is called a *time-homogeneous Markov process*. When no specification is given, a Markov process is assumed first order, time-homogeneous.

III. DERIVATION OF FIRST ORDER MARKOV PROCESS WITH LINEAR CONSTRAINTS

Here we show how the first order Markov process is derived from path entropy maximization. We begin with the definition of path entropy

$$H = -\sum_{\{i_0, i_1, \ldots, i_T\}} p(i_0, i_1, \ldots, i_T) \log p(i_0, i_1, \ldots, i_T). \tag{5}$$

We consider N_1 and N_2 linear constraints on one and two-point probabilities, respectively:

$$\begin{aligned} F_0^{(\alpha)} &\equiv \sum_{t=0}^{T} \sum_{i_t} \varepsilon_{i_t}^{(\alpha)} p(i_t; t) - (T+1) E_0^{(\alpha)} \\ &= 0 \quad (\alpha = 1, \ldots, N_1), \\ F_1^{(\gamma)} &\equiv \sum_{t=0}^{T-1} \sum_{i_t i_{t+1}} J_{i_t i_{t+1}}^{(\gamma)} p(i_t, i_{t+1}; t+1) - T J_0^{(\gamma)} \\ &= 0, \quad (\gamma = 1, \ldots, N_2) \end{aligned} \tag{6}$$

and a normalization condition

$$\sum_{\{i_0, i_1, \ldots, i_T\}} p(i_0, i_1, \ldots, i_T) = 1. \tag{7}$$

These constraints are imposed using Lagrange multipliers. That is, the Lagrange multiplier terms are added to the path entropy as follows:

$$\begin{aligned} &-\sum_{\{i_0, i_1, \ldots, i_T\}} p(i_0, i_1, \ldots, i_T) \log p(i_0, i_1, \ldots, i_T) \\ &-\sum_{\alpha=1}^{N_1} \beta_\alpha \left(\sum_{t=0}^{T} \sum_{i_t} \varepsilon_{i_t}^{(\alpha)} p(i_t; t) - (T+1) E_0^{(\alpha)} \right) \\ &+\sum_{\gamma=1}^{N_2} \nu_\gamma \left(\sum_{t=0}^{T-1} \sum_{i_t i_{t+1}} J_{i_t i_{t+1}}^{(\gamma)} p(i_t, i_{t+1}; t+1) - T J_0^{(\gamma)} \right) \\ &+(\rho+1) \left(\sum_{\{i_0, i_1, \ldots, i_T\}} p(i_0, i_1, \ldots, i_T) - 1 \right). \end{aligned} \tag{8}$$

Extremizing Eq. (8) with respect to $p(i_0, i_1, \ldots, i_T)$, we obtain

$$\begin{aligned} &-\log p(i_0, i_1, \ldots, i_T) - \sum_\alpha \beta_\alpha \sum_{t=0}^{T} \varepsilon_{i_t}^{(\alpha)} \\ &+\sum_\gamma \nu_\gamma \sum_{t=0}^{T-1} J_{i_t i_{t+1}}^{(\gamma)} + \rho = 0. \end{aligned} \tag{9}$$

The Lagrange multipliers introduced in Eq. (8) are determined by additional equations which come from taking the variation of Eq. (8) with respect to these Lagrange multipliers. The solution to Eq. (9) is expressed in terms of the Lagrange multipliers as follows

$$\begin{aligned} &p(i_0, i_1, \ldots, i_T) \\ &= \exp\left(\rho - \sum_\alpha \beta_\alpha \sum_{t=0}^{T} \varepsilon_{i_t}^{(\alpha)} + \sum_\gamma \nu_\gamma \sum_{t=0}^{T-1} J_{i_t i_{t+1}}^{(\gamma)} \right) \\ &= \exp(\rho) v(i_0) G(i_0, i_1) G(i_1, i_2) \ldots G(i_{T-1}, i_T) v(i_T), \end{aligned} \tag{10}$$

where the elements of the vector $\mathbf{v}$, $v(i)$, and the elements of the transfer matrix $\mathbf{G}$, $G(i, j)$, are defined as follows

$$\begin{aligned} v(i) &= \exp\left(-\sum_\alpha \beta_\alpha \varepsilon_i^{(\alpha)} / 2 \right), \\ G(i, j) &= \exp\left(-\sum_\alpha \beta_\alpha \varepsilon_i^{(\alpha)} / 2 + \sum_\gamma \nu_\gamma J_{ij}^{(\gamma)} - \sum_\alpha \beta_\alpha \varepsilon_j^{(\alpha)} / 2 \right). \end{aligned} \tag{11}$$

The m-point joint probability distribution, Eq. (2), is obtained from Eq. (10) by summing over indices $i_0, \ldots, i_{t-m}, i_{t+1}, \ldots, i_T$ as follows:

$$p(a_1,\ldots,a_m;t) = \sum_{i_0,\ldots,i_{t-m},i_{t+1},\ldots,i_T} p(i_0,i_1,\ldots,i_{t-m},a_1,\ldots,a_m,i_{t+1},\ldots,i_T)$$
$$= \exp(\rho)[\mathbf{v}^\dagger\mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)G(a_2,a_3)\ldots G(a_{m-1},a_m)[\mathbf{G}^{T-t}\mathbf{v}](a_m)$$
$$= \frac{[\mathbf{v}^\dagger\mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)G(a_2,a_3)\ldots G(a_{m-1},a_m)[\mathbf{G}^{T-t}\mathbf{v}](a_m)}{\mathbf{v}^\dagger\mathbf{G}^T\mathbf{v}}, \quad (12)$$

where $[\mathbf{v}^\dagger\mathbf{G}^n](a)$ and $[\mathbf{G}^n\mathbf{v}](a)$ denote the ath components of the row and column vectors $\mathbf{v}^\dagger\mathbf{G}^n$ and $\mathbf{G}^n\mathbf{v}$, respectively. (Similarly, $[\mathbf{G}^n](a,b)$ denotes the (a,b) component of the matrix $\mathbf{G}^n$ throughout the paper.) Therefore combining Eqs. (4) and (12), we have

$$p(a_1,\ldots,a_m\to a_{m+1};t) = \frac{\exp(\rho)[\mathbf{v}^\dagger\mathbf{G}^{t-m}](a_1)G(a_1,a_2)\ldots G(a_m,a_{m+1})[\mathbf{G}^{T-t}\mathbf{v}](a_{m+1})}{\exp(\rho)[\mathbf{v}^\dagger\mathbf{G}^{t-m}](a_1)G(a_1,a_2)\ldots G(a_{m-1},a_m)[\mathbf{G}^{T-t+1}\mathbf{v}](a_m)}$$
$$= \frac{G(a_m,a_{m+1})[\mathbf{G}^{T-t}\mathbf{v}](a_{m+1})}{[\mathbf{G}^{T-t+1}\mathbf{v}](a_m)} = p(a_m\to a_{m+1};t). \quad (13)$$

This shows that a conditional probability of transition in fact depends only on the last two states, those right before and after the transition. Therefore, the process is indeed a first order Markov one. However, it should be noted that the transition probability has explicit time dependence.

The first order Markov property was also derived in Ref. 25 for the special case of constraining one-point and two-point statistics which we now define. One-particle statistics corresponds to $F_0^{(\alpha)}$ with

$$\varepsilon_i^{(\alpha)} = \delta_{i,\alpha} \quad (\alpha = 1,\ldots,N), \quad (14)$$

where the index α of the constraint now goes over each state of the system, N being their total number of such states. This constraint simply counts the number of times state α is visited over the course of the trajectory. Likewise, two-point statistics corresponds to imposing $F_1^{(\tau,\sigma)}$ with

$$J_{i,j}^{(\tau,\sigma)} = \delta_{i,\tau}\delta_{j,\sigma} \quad (\tau,\sigma = 1,\ldots,N), \quad (15)$$

where we labelled the constraint by double indices (τ,σ) instead of the single index γ for notational convenience. This again simply counts the number of transitions from state τ to σ over the course of the trajectory.

IV. DERIVATION OF THE TIME-HOMOGENEOUS MASTER EQUATION

Recall that a master equation requires time-dependent state occupation probabilities and time-independent transition probabilities. Under what conditions are such approximations valid? To answer this question we apply the Perron-Frobenius theorem[26–29] to the $\mathbf{G}$ transfer matrix of Sec. III—a square matrix which by construction is of size $N\times N$ and has positive elements. According to the theorem, $\mathbf{G}$ satisfies the following properties:

(1) It has a positive real eigenvalue r, called the Perron-Frobenius eigenvalue, such that any other eigenvalue λ is strictly smaller than r in absolute value, $|\lambda| < r$.
(2) There is a left eigenvector $\mathbf{y}^\dagger = (y_1,\ldots,y_N)$ for r with positive components. That is, $\mathbf{y}^\dagger\mathbf{G} = r\mathbf{y}^\dagger$ and $y_i > 0$ for all i. Similarly, there is a right eigenvector $\mathbf{z}$ with positive components, such that $\mathbf{G}\mathbf{z} = r\mathbf{z}$ and $z_i > 0$ for all i.
(3) Left and right eigenvectors with eigenvalue r are non-degenerate.
(4) $\lim_{T\to\infty}\frac{\mathbf{G}^T}{r^T} = \mathbf{z}\mathbf{y}^\dagger$

Now reconsider Eq. (13) where

$$p(a_m\to a_{m+1};t) = \frac{G(a_m,a_{m+1})[\mathbf{G}^{T-t}\mathbf{v}](a_{m+1})}{[\mathbf{G}^{T-t+1}\mathbf{v}](a_m)}. \quad (16)$$

Since the vector $\mathbf{v}$ has only non-negative elements, both $\mathbf{G}^\mathbf{T}\mathbf{v}/r^T$ and $\mathbf{v}^\dagger\mathbf{G}^\mathbf{T}/r^T$ have well-defined non-zero limits for $T\to\infty$,

$$\lim_{T\to\infty}\frac{\mathbf{G}^T\mathbf{v}}{r^T} = \mathbf{z}(\mathbf{y}^\dagger\mathbf{v}); \quad \lim_{T\to\infty}\frac{\mathbf{v}^\dagger\mathbf{G}^T}{r^T} = (\mathbf{v}^\dagger\mathbf{z})\mathbf{y}^\dagger. \quad (17)$$

Therefore, taking the limit $T-t\to\infty$ of Eq. (16) and using Eq. (17), we find

$$p(a\to b) = \frac{G(a,b)z(b)}{rz(a)}. \quad (18)$$

That is, the transition probability is time-independent in this limit. However, from Eq. (12), the m-point joint probabilities are still explicitly time-dependent when $T-t$ is large

$$p(a_1,\ldots,a_m;t) = \frac{[\mathbf{v}^\dagger\mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)G(a_2,a_3)\ldots G(a_{m-1},a_m)z(a_m)}{r^t\mathbf{v}^\dagger\mathbf{z}} \quad (19)$$

and, in particular, this is true for the one-point occupation probability

$$p(a;t) = \frac{[\mathbf{v}^\dagger\mathbf{G}^t](a)z(a)}{r^t\mathbf{v}^\dagger\mathbf{z}}. \quad (20)$$

Thus maximizing the path entropy under the linear constraint Eq. (6) up to two-point probabilities, which are imposed for infinite duration into the future ($T-t\to\infty$), we obtain a time-homogeneous Markov process which is described by (1) time-independent transition probabilities and (2) time-dependent one-point occupation probabilities. From

Eqs. (18) and (20), we now obtain the evolution equation for the time-homogeneous Markov process

$$p(a;t+1) = \sum_b p(b;t)p(b \to a), \tag{21}$$

which is the celebrated master equation.

Note the asymmetry in time: the transition probabilities as well as the joint probabilities are time dependent when the limit of $t \to \infty$ is taken but $T - t$ is kept finite. This is simply due to the fact that the transition probability $p(b \to a)$ is defined in a time-asymmetric manner.

The last limit to consider is the stationary case, when both $T - t$ and t are large. Then the m-point joint probability of Eq. (12) reduces to

$$p(a_1, \ldots, a_m) = \frac{y(a_1)G(a_1,a_2)G(a_2,a_3)\ldots G(a_{m-1},a_m)z(a_m)}{r^{m-1}\mathbf{y}^\dagger\mathbf{z}}, \tag{22}$$

which is independent of time as are the state occupation probability or any conditional probability derived from Eq. (22). This is to be expected, since we have time translation invariance in the stationary limit. Stationarity also trivially follows when the constraints themselves are stationary, which are much stronger conditions than those in Eq. (6).

Equation (22) was also derived in the large T limit with $(m = T)$ in Ref. 31 though the stationary Markov process was assumed from the onset therein. In contrast, in the current work m is finite and can be as small as 1, even in the large T limit, and stationarity is derived rather than being an *a priori* assumption. Likewise, the first order Markov process was derived in Ref. 25 from path entropy maximization for the special case of pair statistics constraints, but neither conditions for the time-homogeneous process nor stationarity were discussed.[32]

V. TIME-HOMOGENEOUS MARKOV PROCESSES WITH AN ARBITRARY INITIAL CONDITION

We have discussed how data can come in the form of state occupation probabilities (e.g., how long during the course of a single molecule fluorescence experiment did a protein dwell in a low fluorescent state) or transition probabilities. However, data may also be available in the form of conditions at different points in time (e.g., the sample is pumped into a photoexcited state at time $t = 0$). Are our conclusions on time-homogeneity from Sec. IV robust to initial, final, or other such conditions? In this section, we briefly show when the time-homogeneity of transition probability depends on such conditions.

Consider an arbitrary condition imposed at time τ

$$p(a;t=\tau) = \pi(a). \tag{23}$$

We then add the term $\sum_a \lambda(a)(p(a;\tau) - \pi(a))$ with Lagrange multipliers $\lambda(a)$ $(a = 1, \ldots, N)$ to the constrained entropy, Eq. (8). As before, setting the variation with respect to $p(i_0, i_1, \ldots, i_T)$ to zero yields

$$\begin{aligned} p(i_0,i_1,\ldots,i_T) &= \exp\left(\rho + \lambda(i_\tau) - \beta\sum_{t=0}^{T}\varepsilon_{i_t} + \nu\sum_{t=0}^{T-1}J_{i_t i_{t+1}}\right) \\ &= \exp(\rho+\lambda(i_\tau))v(i_0)G(i_0,i_1)G(i_1,i_2)\ldots G(i_{T-1},i_T)v(i_T) \\ &= \frac{v(i_0)\pi(i_\tau)G(i_0,i_1)G(i_1,i_2)\ldots G(i_{T-1},i_T)v(i_T)}{\sum_{j_0\ldots,j_T} v(j_0)\pi(j_\tau)G(j_0,j_1)G(j_1,j_2)\ldots G(j_{T-1},j_T)v(j_T)}, \end{aligned} \tag{24}$$

where in the last line we used the normalization condition Eq. (7) to eliminate ρ and the initialization constraint Eq. (23) to eliminate λ. We now have

$$\begin{aligned} &\tau \le t - m + 1: \\ &p(a_1,\ldots,a_m;t) = \frac{\sum_a [\mathbf{v}^\dagger\mathbf{G}^\tau](a)\pi(a)[\mathbf{G}^{t-\tau-m+1}](a,a_1)G(a_1,a_2)\ldots G(a_{m-1},a_m)[\mathbf{G}^{T-t}\mathbf{v}](a_m)}{\sum_b [\mathbf{v}^\dagger\mathbf{G}^\tau](b)\pi(b)[\mathbf{G}^{T-\tau}\mathbf{v}](b)}, \\ &t - m + 1 < \tau \le t: \\ &p(a_1,\ldots,a_m;t) \\ &\quad = \frac{[\mathbf{v}^\dagger\mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)\ldots G(a_{\tau-t+m-1},a_{\tau-t+m})\pi(a_{\tau-t+m})}{\sum_b [\mathbf{v}^\dagger\mathbf{G}^\tau](b)\pi(b)[\mathbf{G}^{T-\tau}\mathbf{v}](b)} \\ &\qquad \times G(a_{\tau-t+m},a_{\tau-t+m+1})\ldots G(a_{m-1},a_m)[\mathbf{G}^{T-t}\mathbf{v}](a_m), \\ &t < \tau: \\ &\quad p(a_1,\ldots,a_m;t) = \frac{\sum_a [\mathbf{v}^\dagger\mathbf{G}^{t-m+1}](a_1)G(a_1,a_2)\ldots G(a_{m-1},a_m)}{\sum_b [\mathbf{v}^\dagger\mathbf{G}^\tau](b)\pi(b)[\mathbf{G}^{T-\tau}\mathbf{v}](b)} \\ &\qquad \times [\mathbf{G}^{\tau-t}](a_m,a)\pi(a)[\mathbf{G}^{T-\tau}\mathbf{v}](a). \end{aligned} \tag{25}$$

Using the definition of the transition probability from Eq. (4) we find

$$\tau < t:$$

$$p(a_1,\ldots,a_m \to a_{m+1};t) = \frac{G(a_m,a_{m+1})[\mathbf{G}^{T-t}\mathbf{v}](a_{m+1})}{[\mathbf{G}^{T-t+1}\mathbf{v}](a_m)}$$

$$\tau \geq t:$$

$$p(a_1,\ldots,a_m \to a_{m+1};t) = \frac{G(a_m,a_{m+1})\sum_a[\mathbf{G}^{\tau-t}](a_{m+1},a)\pi(a)[\mathbf{G}^{T-\tau}\mathbf{v}](a)}{\sum_b[\mathbf{G}^{\tau-t+1}](a_m,b)\pi(b)[\mathbf{G}^{T-\tau}\mathbf{v}](b)}. \quad (26)$$

We notice that the indices $a_1,\ldots,a_{m-1}$ have dropped out from the right-hand side of Eq. (26). We can therefore write

$$p(a_1,\ldots,a_m \to a_{m+1};t) = p(a_m \to a_{m+1};t), \quad (27)$$

showing that, once more, we have a first order Markov process. Furthermore, the transition probability for $t > \tau$ has exactly the same form as Eq. (13), independent of the initial condition $\boldsymbol{\pi}$. It is therefore time-homogeneous under the limit of large $T - t$. The same is not true of $t \leq \tau$, where the transition probability always depends on the specified condition and time-homogeneity requires both large $T - \tau$ and $\tau - t$. As noted earlier, this time-asymmetry is a natural consequence of the fact that the definition of the transition probability itself is time-asymmetric.

VI. GENERAL DERIVATION OF *n*th-ORDER MARKOV PROCESS FROM PATH ENTROPY MAXIMIZATION

In this section, we generalize the arguments of Sec. III in two important ways: (1) we consider constraints on the data up to $n+1$-point probabilities

$$F^{(\alpha)}(\{p(i;t)\},\{p(i\to j;t)\},\ldots, \{p(i_0,\ldots,i_{n-1}\to i_n;t)\}) = 0, \quad (28)$$

and (2) we do not assume that the constraints $F^{(\alpha)}$ are linear functions of their arguments (as was the case for Eq. (6)).

Provided constraints are linear—as was the case in Eq. (6)—the arguments in Sec. III are generalizable to *n*th-order Markov processes. Indeed, the path probability would be described by the multiplication of rank-$(n+1)$ tensors rather than matrices, such as Eq. (12). The *n*th-order Markov process would follow immediately, though the derivation of the time-homogeneity of various transition probabilities, as in Secs. IV and V, would require the difficult task of applying an analogue of the Perron-Frobenius theorem for general tensors.

Since we want to derive the *n*th-order Markov process for fully general constraints, as given by Eq. (28), we take a different route. We first express the path probability $p(i_1, i_2,\ldots,i_T)$ in terms of the conditional probabilities:

$$p(i_0,i_1,\ldots,i_T) = p(i_0;0)p(i_0\to i_1;1)p(i_0,i_1\to i_2;2)\ldots \times p(i_0,i_1\ldots,i_{T-1}\to i_T;T). \quad (29)$$

Substituting this expression into Eq. (5), we get

$$\begin{aligned} H &= -\sum_{\{i_0,i_1,\ldots,i_T\}} p(i_0,i_1,\ldots,i_T) \\ &\quad\times\left(\log p(i_0;0) + \sum_{t=0}^{T-1}\log p(i_0,\ldots,i_t\to i_{t+1};t+1)\right) \\ &= -\sum_i p(i;0)\log p(i;0) \\ &\quad -\sum_{t=0}^{T-1}\sum_{\{i_0,i_1,\ldots,i_{t+1}\}} p(i_0,i_1,\ldots,i_{t+1};t+1) \\ &\quad\times\log p(i_0,\ldots,i_t\to i_{t+1};t+1), \end{aligned} \quad (30)$$

where, in getting from first to second line, we invoked the relation between joint and marginal probabilities; $p(i_0,\ldots,i_m;m) = \sum_{i_{m+1},\ldots,i_T} p(i_1,i_2,\ldots,i_T)$.

Now reconsider the constraints given by Eq. (28) imposed from $p(i;t)$ to $p(i_0,\ldots,i_{n-1}\to i_n;t)$. We will maximize the entropy, Eq. (30), in two steps:

(1) We maximize the entropy with respect to $\{p(i_0,\ldots,i_k;t)\}$ $(k > n)$, for given values of $\{p(i_0,\ldots,i_k;t)\}$ with $k \leq n$.
(2) We then vary the entropy over the remaining variables, $\{p(i_0,\ldots,i_k;t)\}$ $(k \leq n)$.

By assumption, constraints on the data only matter in step 2. Furthermore, as we now show, step 1 (the unconstrained maximization) is sufficient to show that the general path probability reduces to that of an *n*th-order Markov process.

In order to perform step 1, we first invoke the equality

$$-\sum_i q_i\log q_i \leq -\sum_i q_i\log p_i \quad (31)$$

for arbitrary probability distributions p_i and q_i.[33] It follows from Eq. (31) that

$$\begin{aligned} &-\sum_j p(i_0,\ldots,i_{m-1}\to j;t)\log p(i_0,\ldots,i_{m-1}\to j;t) \\ &\leq -\sum_j p(i_0,\ldots,i_{m-1}\to j;t)\log p(i_{m-n},\ldots,i_{m-1}\to j;t). \end{aligned} \quad (32)$$

Multiplying both sides of Eq. (32) by $p(i_0,\ldots,i_{m-1};t)$ and summing over $i_0,\ldots,i_{m-1}$, we find

$$\begin{aligned} &= -\sum_{i_0,\ldots,i_{m-1},j} p(i_0,\ldots,i_{m-1},j;t)\log p(i_0,\ldots,i_{m-1}\to j;t) \\ &\leq -\sum_{i_0,\ldots,i_{m-1},j} p(i_0,\ldots,i_{m-1},j;t)\log p(i_{m-n},\ldots,i_{m-1}\to j;t), \end{aligned} \quad (33)$$

where Eq. (4) was used. The above sets a bound on the last term of the path entropy, Eq. (30). Therefore, for given values of $\{p(i_0,\ldots,i_k;t)\}$ with $k \leq n$, we see that H is maximized for

$$p(i_0,\ldots,i_{m-1}\to j;t) = p(i_{m-n},\ldots,i_{m-1}\to j;t) \quad (m > n), \quad (34)$$

the system now being described by a nth-order Markov model where the probability $p(i; t)$ is determined only by previous n steps of history.

Now the transition probability Eq. (34) for nth-order Markov process can be substituted into the path entropy formula, Eq. (30). Step 2 can then be carried forward: the resulting path entropy can be maximized with respect to the remaining variables $p(i_0; t)$, $p(i_0 \to i_1; t), \ldots, p(i_0, i_1, \ldots, i_{n-1} \to i_n; t)$ under the constraints, Eq. (28).

In summary, we have just shown that nth-order Markov processes follow under very general constraints provided by Eq. (28). Markov models emerge from the entropy maximization method—and these provide immediate and principled generalizations of the ubiquitous master equation.

VII. DISCUSSION

Markov processes and master equations—the evolution equation describing a first order time-homogeneous Markov process—are standard stochastic modeling tools invoked across disciplines. Such models are usually justified mechanistically by coarse-graining arguments or by assuming quick randomization in space of reactants and products (the "well-stirred" approximation). Yet it is challenging to ascertain *a priori* whether any of these conditions actually hold. Just like maximum entropy has provided an alternative to ergodic theory for the justification of the equilibrium probability distribution,[17] we believe that the path entropy techniques of Filyukov and Karpov,[14] and later Jaynes,[15] provide a compelling axiomatic basis for the Markov process and the master equation. Here the Markov process emerges as a solution to the following inverse problem: given measurable n-point constraints on a trajectory, what is the least biased model for a probability distribution of paths? By least biased, we mean one that, for instance, does not impose correlations in a model when such correlations are not otherwise warranted by the data (technically these are the logical consistency axioms of Shore and Johnson). The unique solution to this problem is that which maximizes the entropy subject to constraints from the data.

With this formalism, we justify generalizations of the master equation on rigorous mathematical grounds. It is tempting to conjecture whether the nth-order Markov process can lead to a time-homogeneous process so long as the constraints are imposed for a time much longer than that of one time step. The proof would require an analogue of the Perron-Frobenius theorem for general tensors, an interesting subject for further investigation.

ACKNOWLEDGMENTS

We thank Ken Dill, Kingshuk Ghosh, and Hao Ge for useful discussions. S.P. acknowledges an FQRNT fellowship and Ken Dill's support by way of NSF Grant No. R01GM090205-03.

[1] N. G. van Kampen, *Stochastic Processes in Chemistry and Physics* (North-Holland, Amsterdam, 1981).
[2] K. L. Chung, *Lectures from Markov Processes to Brownian Motion* (Springer-Verlag, New York, 1982).
[3] I. Gopich and A. Szabo, J. Chem. Phys. **118**, 454 (2003).
[4] J. Cao and R. J. Silbey, J. Phys. Chem. B **112**, 12867 (2008).
[5] A. M. Berezhkovskii, A. Szabo, and G. H. Weiss, J. Chem. Phys. **110**, 9145 (1999).
[6] X.-J. Zhang, H. Qian, and M. Qian, Phys. Rep. **510**, 1 (2012).
[7] H. Ge, H. Qian, and M. Qian, Phys. Rep. **510**, 87 (2012).
[8] F. L. H. Brown, Acc. Chem. Res. **39**, 363 (2006).
[9] H. D. Feng and J. Wang, Chem. Phys. Lett. **501**, 562 (2011).
[10] S. R. de Groot and P. Mazur, *Non-Equilibrium Thermodynamics* (Dover, New York, 1983).
[11] J.-H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith, and F. Noé, J. Chem. Phys. **134**, 244108 (2011).
[12] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99 (2010).
[13] P. Kasson and V. S. Pande, Pac. Symp. Biocomput. **15**, 260 (2010).
[14] A. A. Filyukov and V. Y. Karpov, J. Eng. Phys. Thermophys. **13**, 326 (1967); **13**, 416 (1967); A. A. Filyukov, *ibid.* **14**, 429 (1968).
[15] E. T. Jaynes, "Macroscopic prediction," in *Complex Systems Operational Approaches in Neurobiology, Physics, and Computers*, edited by H. Haken (Springer-Verlag, Berlin, 1985).
[16] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **26**, 26 (1980).
[17] E. T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).
[18] S. F. Gull and G. J. Daniell, Nature (London) **272**, 686 (1978).
[19] P. J. Steinbach, K. Chu, H. Frauenfelder, J. B. Johnson, D. C. Lamb, G. U. Nienhaus, T. B. Sauke, and R. D. Young, Biophys. J. **61**, 235 (1992).
[20] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, London, 2003).
[21] G. Stock, K. Ghosh, and K. A. Dill, J. Chem. Phys. **128**, 194102 (2008).
[22] K. Ghosh, K. A. Dill, M. M. Inamdar, E. Seitaridou, and R. Phillips, Am. J. Phys. **74**, 123 (2006).
[23] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, New York, 2001).
[24] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[25] H. Ge, S. Pressé, K. Ghosh, and K. A. Dill, J. Chem. Phys. **136**, 064108 (2012).
[26] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences* (SIAM, 1994).
[27] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability* (Springer-Verlag, London, 1993).
[28] E. Seneta, *Non-Negative Matrices and Markov Chains* (Springer, 1981).
[29] D. L. Isaacson and I. Madsen, *Markov Chains: Theory and Applications* (Wiley, 1976).

[30] The stationary process is also obtained when the constraint are imposed at each point in time:

$$\begin{aligned} F_0^{(\alpha)}(t) &= \varepsilon_{i_t}^{(\alpha)} p(i_t; t) - E_0^{(\alpha)} \\ &= 0 \quad (\alpha = 1, \ldots, N_1)(t = 0, \ldots, T), \\ F_1^{(\gamma)}(t) &= \sum_{i_t i_{t+1}} J_{i_t i_{t+1}}^{(\gamma)} p(i_t, i_{t+1}; t+1) - J_0^{(\gamma)} \\ &= 0 \quad (\gamma = 1, \ldots, N_2)(t = 0, \ldots, T-1). \end{aligned}$$

Our result shows that the weaker constraint Eq. (6) can achieve this so long as $0 \ll t, T - t$.

[31] C. J. Monthus, J. Stat. Mech.: Theory Exp. **2011**, P03008.

[32] Adapted to our notation, it is stated underneath of Eq. (11) of Ref. 25, that $p(a, b) \propto G(a, b)$, implying that $p(a, b)$ is time-independent. However, since $p(a, b) = \frac{[\mathbf{v}^\dagger \mathbf{G}^{t-1}](a) G(a,b) [\mathbf{G}^{T-t}\mathbf{v}](b)}{\mathbf{v}^\dagger \mathbf{G}^T \mathbf{v}}$ from Eq. (12), this is strictly correct only when $T - t$ and t are both large.

[33] Using the well-known inequality $\log x \le -1 + x$ for $x > 0$, we see that

$$\begin{aligned} &-\sum q_i \log q_i + \sum q_i \log p_i \\ &= \sum q_i \log \frac{p_i}{q_i} \le \sum_i q_i \left(-1 + \frac{p_i}{q_i}\right) \\ &= -\sum_i q_i + \sum_i p_i = 0, \end{aligned}$$

proving the inequality, Eq. (31). This inequality was also invoked in Ref. 14 in a much narrower setting (of deriving a 0th order Markov model).