

# Structure and Randomness of Continuous-Time, Discrete-Event Processes

Sarah E. Marzen<sup>1,2</sup> · James P. Crutchfield<sup>3</sup>

Received: 15 April 2017 / Accepted: 8 August 2017 / Published online: 30 August 2017  
© Springer Science+Business Media, LLC 2017

**Abstract** Loosely speaking, the Shannon entropy rate is used to gauge a stochastic process' intrinsic randomness; the statistical complexity gives the cost of predicting the process. We calculate, for the first time, the entropy rate and statistical complexity of stochastic processes generated by finite unifilar hidden semi-Markov models—memoryful, state-dependent versions of renewal processes. Calculating these quantities requires introducing novel mathematical objects ( $\epsilon$ -machines of hidden semi-Markov processes) and new information-theoretic methods to stochastic processes.

**Keywords** Epsilon-machines · Causal states · Entropy rate · Statistical complexity · Hidden Markov processes

## 1 Introduction

Claude Shannon's seminal 1948 article "A Mathematical Theory of Communication" introduced a definition for entropy as a well-motivated measure of randomness [1]. He further identified entropy rate  $h_\mu$  as a measure of the minimal coding cost of a series of potentially correlated symbols in his celebrated first theorem. In 1989, Young and Crutchfield identified statistical complexity  $C_\mu$  as the entropy of causal states [2], which are the minimal sufficient statistics of prediction [3]. Said simply,  $h_\mu$  is a measure of a process' intrinsic random-

---

✉ James P. Crutchfield  
chaos@ucdavis.edu

Sarah E. Marzen  
semarzen@mit.edu

<sup>1</sup> Physics of Living Systems Group, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

<sup>2</sup> Department of Physics, University of California, Berkeley, Berkeley, CA 94720-5800, USA

<sup>3</sup> Complexity Sciences Center, Department of Physics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

ness and  $C_\mu$  a measure of process structure. Both entropy rate and statistical complexity have given insight into many disparate complex systems, from chaotic crystallography [4], biomolecule dynamics [5–7], neural spike trains [8], and animal behavior [9] to stochastic resonance [10], geomagnetic volatility [11], hydrodynamic flows [12, 13], and fluid and atmospheric turbulence [14, 15].

These two measures of complexity are unified by the Kolmogorov–Chaitin complexity of a discrete object, which is the size of the minimal Universal Turing Machine program that produces the object [16, 17]. Specifically, the expected Kolmogorov–Chaitin complexity ( $K(x_\ell)$ ) of a (discrete-time, discrete-symbol) time series  $x_\ell$  of length  $\ell$  grows at the Shannon entropy rate and has an offset determined by the statistical complexity:  $\log(K(x_\ell)) \propto_{\ell \rightarrow \infty} C_\mu + \ell h_\mu$ , when these quantities exist [18].

Perhaps somewhat surprisingly, estimators of the entropy rate and statistical complexity of *continuous-time*, discrete-symbol processes are lacking. This is unfortunate since these processes are encountered very often in the physical, chemical, biological, and social sciences as sequences of discrete events consisting of an event type and an event duration or magnitude. An example critical to infrastructure design occurs in the geophysics of crustal plate tectonics, where the event types are major earthquakes tagged with duration time, time between their occurrence, and an approximate or continuous *Richter magnitude* [19]. Understanding these process' randomness and structure bears directly on averting human suffering. Another example is revealed in the history of reversals of the earth's geomagnetic field [11], which shields the planet's life from exposure to damaging radiation. An example from physical chemistry is found in single-molecule spectroscopy which reveals molecular dynamics as hops between conformational states that persist for randomly distributed durations [6, 7]. The structure of these conformational transitions is implicated in biomolecular functioning and so key to life processes. A common example from neuroscience is found in the spike trains generated by neurons that consist of spike-no-spike event types separated by continuous *interspike intervals*. The structure and randomness of spike trains are key to delineating how tissues support essential information processing in nervous systems and brains [8]. Finally, a growing set of these processes appear in the newly revitalized quantitative social sciences, in which human communication events and their durations are monitored as signals of emergent coordination or competition [20].

Here, we provide new closed-form expressions for the entropy rate and statistical complexity of continuous-time, discrete-symbol processes by first identifying the causal states of stochastic processes generated by continuous-time unifilar hidden semi-Markov models, a restricted class of the models described in Ref. [21]. In these processes successive dwell times for the symbols (or events) are drawn based on the process' current "hidden" state. Transitions from hidden state to hidden state then follow a rule that mimics transitions in discrete-time unifilar hidden Markov models. The resulting output process consists of the symbols emitted during the dwell time. Identifying the process causal states leads to new expressions for entropy rate, generalizing one of the results given in Sect. 7.4 of Ref. [22], and for statistical complexity. The hidden semi-Markov process class we analyze here is sufficiently general we conjecture that our results yield universal estimators of the entropy rate and statistical complexity of continuous-time, discrete-event processes.

To start, we define unifilar hidden semi-Markov processes and their generators, determine their causal states, and use the causal states to calculate their entropy rate and statistical complexity. We conclude by describing a method for using the expressions given here to

estimate the entropy rate and statistical complexity of continuous-time, discrete-event time series.

## 2 Hidden Semi-Markov Processes and Their Unifilar Generators

The continuous-time, discrete-symbol process  $\dots, (X_{-1}, T_{-1}), (X_0, T_0), (X_1, T_1), \dots$  has realizations  $\dots (x_{-1}, \tau_{-1}), (x_0, \tau_0), (x_1, \tau_1) \dots$ . Events are symbol-duration pairs  $(x_i, \tau_i)$  that occur sequentially in a process. For symbols, we demand that  $x_i \neq x_{i+1}$  to enforce a unique description of the process. In other words, the discrete event symbol  $x_i \in \mathcal{A}$  appears for a total time of  $\tau_i$ . The present is located almost surely during the emission of  $x_0$ , and we denote the time since last emission as  $\tau_{0+}$  and the time to next emission as  $\tau_{0-}$ . We also, for clarity, denote the last-appearing symbol as  $x_{0+}$  and the next-appearing symbol as  $x_{0-}$ ; though obviously  $x_{0+} = x_{0-}$  almost surely. (It follows that  $\tau_{0+} + \tau_{0-} = \tau_0$ .) The *past*  $\dots, (X_{-1}, T_{-1}), (X_0, T_{0+})$  is denoted  $\overleftarrow{(X, T)}$  and the *future*  $(X_0, T_{0-}), (X_1, T_1), \dots$  is denoted as  $\overrightarrow{(X, T)}$ , with realizations denoted  $\overleftarrow{(x, \tau)}$  and  $\overrightarrow{(x, \tau)}$ , respectively.

A continuous-time, discrete-symbol process' causal states  $\mathcal{S}$  are the equivalence classes of pasts defined by the relation:

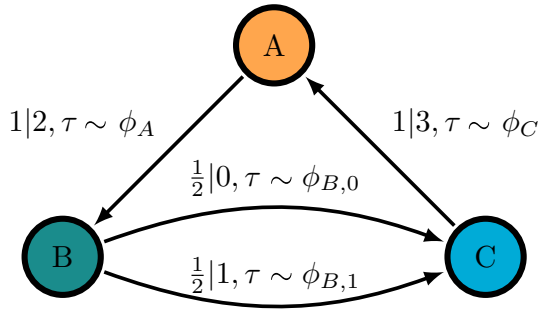
$$\overleftarrow{(x, \tau)} \sim_\epsilon \overleftarrow{(x', \tau')} \iff \Pr\left(\overrightarrow{(X, T)} \mid \overleftarrow{(X, T)} = \overleftarrow{(x, \tau)}\right) = \Pr\left(\overrightarrow{(X, T)} \mid \overleftarrow{(X, T)} = \overleftarrow{(x', \tau')}\right). \tag{1}$$

This mimics the relation for the causal states of discrete-time processes [3]. A process' *statistical complexity* is the entropy of these causal states:  $C_\mu = H[\mathcal{S}]$ . A process' *precise states* are any finer-grained partition of the causal-state classes, as for discrete-time processes [3]. Thus, there is a fundamental distinction between a process' observed or emitted symbols and its internal states. In this way, we consider general processes as *hidden* processes.

Causal states can be discrete random variables, continuous random variables, or mixed random variables. For discrete-event, continuous-time processes, the last of these is the likeliest. The entropy of discrete random variables and the differential entropy of continuous random variables are both given in Ref. [23], while the entropy of mixed random variables is described in Ref. [24]; all are represented as  $H[X]$ , where  $X$  is the random variable. Conditional entropy  $H[X|Y]$  is then defined as  $\langle H[X|Y = y] \rangle_y$ . The mutual information between random variables  $X$  and  $Y$  is given by  $I[X; Y] = H[X] - H[X|Y]$ , and numerous other definitions are given in Ref. [23].

A *hidden semi-Markov process* (HSMP) is a continuous-time, discrete-symbol process generated by a *hidden semi-Markov model* (HSMM). A HSMM is described via a hidden-state random variable  $\mathcal{G}$  with realization  $g$ , an emission probability  $T_g^{(x)}$ , which is the probability of emitting  $x$  when in hidden state  $g$ , and a dwell-time distribution  $\phi_g(\tau)$ . In other words, in hidden state  $g$ , symbol  $x$  is emitted with probability  $T_g^{(x)}$  for time  $\tau$  drawn from  $\phi_g(\tau)$ . For reasons that will become clear, we focus on a restricted form—the *unifilar* HSMM (uHSMM). For these, the present hidden state  $g_0$  is uniquely determined by the past emitted symbols  $x_{-\infty:0} = \dots, x_{-2}, x_{-1}$ . See Fig. 1. In an abuse of notation, Eq. (1) determines a function  $\epsilon(x_{-\infty:0})$  that takes the past emitted symbol sequence  $x_{-\infty:0}$  to the underlying hidden state  $g_0$ . (The abuse comes from suppressing the dependence on durations.) This is the analog appropriate to this setting of the familiar definition of unifilarity in discrete-time models [3]—that symbol and state uniquely determine next state.

**Fig. 1** Generative model for a unifilar hidden semi-Markov process. The notation  $p(x)|x, \tau \sim \phi_g$  means that  $x$  is emitted with probability  $p(x)$  from hidden state  $g$  with emission time drawn from  $\phi_g$



### 3 Causal Architecture

The challenge now is to identify a given HSMP’s causal states. With the causal states in hand, we can calculate their statistical complexity and entropy rate rather straightforwardly. Theorem 1 makes the identification.

**Theorem 1** *A unifilar hidden semi-Markov process’ causal states are the triple  $(g_0, x_{0+}, \tau_{0+})$ , under weak assumptions specified below.*

*Proof* To identify causal states, we need to simplify the conditional probability distribution:

$$p\left(\overrightarrow{(x, \tau)} \mid \overleftarrow{(x, \tau)}\right) = \Pr\left(\overrightarrow{(X, T)} = \overrightarrow{(x, \tau)} \mid \overleftarrow{(X, T)} = \overleftarrow{(x, \tau)}\right).$$

When the process is generated by a unifilar hidden semi-Markov model, this conditional probability distribution simplifies to:

$$\begin{aligned} p\left(\overrightarrow{(x, \tau)} \mid \overleftarrow{(x, \tau)}\right) &= p(x_{0-}, \tau_{0-}, (x_1, \tau_1), \dots \mid \dots, (x_{-1}, \tau_{-1}), (x_{0+}, \tau_{0+})) \\ &= p((x_1, \tau_1), \dots \mid \dots, (x_0, \tau_0)) p(\tau_{0-} \mid \dots, (x_{-1}, \tau_{-1}), (x_{0+}, \tau_{0+}), x_{0-}) \\ &\quad \times p(x_{0-} \mid \dots, (x_{-1}, \tau_{-1}), (x_{0+}, \tau_{0+})), \end{aligned} \tag{2}$$

where we have  $\tau_0 = \tau_{0+} + \tau_{0-}$ . Almost surely, we have:

$$p(x_{0-} \mid \dots, (x_{-1}, \tau_{-1}), (x_{0+}, \tau_{0+})) = \delta_{x_{0+}, x_{0-}} \tag{3}$$

and, due to the unifilarity constraint:

$$p(\tau_{0-} \mid \dots, (x_{-1}, \tau_{-1}), (x_{0+}, \tau_{0+}), x_{0-}) = p(\tau_{0-} \mid \tau_{0+}, g_0 = \epsilon^+(x_{0-})). \tag{4}$$

Together Eqs. (2)–(4) imply that the triple  $(g_0, x_{0+}, \tau_{0+})$  are prescient statistics.

These states are causal (minimal and prescient) when three things happen: first, when the number of hidden states  $g$  is the smallest possible; second, when  $\phi_g(\tau)$  does not take either the “eventually Poisson” or “eventually  $\Delta$ -Poisson” form described in Ref. [25]; and third, when two of the “conveyor belts” as shown in Fig. 3 do not merge.

Each of these conditions is worth spelling out.

The first condition states that of all such unifilar hidden semi-Markov models which generate the given process, we choose the model with the smallest number of hidden states. Two hidden states can be considered equivalent when they have identical future morphs,  $p(\overrightarrow{(x, \tau)} \mid g, x_+, \tau_+)$ . Here, Eqs. (2)–(4) aid in detecting such identical hidden states.

As for the second condition, to avoid an eventually Poisson-like dwell time distribution, we demand that  $\phi_g(\tau)$  cannot be written as  $\phi_g(T)e^{-\lambda(t-T)}$ , for all  $t \geq T$  and some  $\lambda > 0$  and  $T \geq 0$ . To avoid an eventually  $\Delta$ -Poisson-like dwell time distribution, we demand that  $\phi_g(\tau)$  cannot be written as:

$$\phi_g(t) = \phi_g(T + (t - T) \bmod \Delta) e^{-\lambda \lfloor (t-T)/\Delta \rfloor},$$

for any  $0 < \Delta, \lambda, T < \infty$ . Almost all naturally occurring dwell-time distributions take neither of these forms, though a Poisson neuron with a refractory period generates an eventually Poisson renewal process.

Alternatively, and this is the third condition, it is possible that  $\epsilon(g_0, x_0) = \epsilon(g'_0, x_0)$  and  $p(\tau_{0-} | g_0, \tau_{0+}) = p(\tau_{0-} | g'_0, \tau'_{0+})$  for all  $\tau_{0-}$ . Here, the notation is overloaded:  $\epsilon(g_0, x_0)$  is the hidden state to which we transition given that we are in hidden state  $g_0$  and observe symbol  $x_0$ . In this case, the conditional probability distributions in Eq. (2) are equal, despite the different hidden states and potentially different times since last event. Such an equality would engender merging of the renewal process-like conveyor belts in Fig. 3. And so, we stipulate that there does not exist a pair  $g_0 \neq g'_0$  and  $\tau_{0+}$  and  $\tau'_{0+}$  such that  $p(\tau_{0-} | g_0, \tau_{0+}) = p(\tau_{0-} | g'_0, \tau'_{0+})$  for all  $\tau_{0-}$ . This condition, described in more detail below, is shown in Fig. 2 and is almost always met for the applications cited earlier. Thus, we say that the typical unifilar hidden semi-Markov process' causal states are given by the triple  $(g_0, x_{0+}, \tau_{0+})$ .

The statistics described in Theorem 1 are still prescient statistics—i.e., sufficient statistics of prediction—even when the weak assumptions are not met. In discrete-time applications, this would imply that the entropy of those prescient states upper-bounds the statistical complexity. In continuous-time applications, prescient states and causal states are often mixed random variables and, so, no such upper bound can be stated generally.

As a result, it is well worth spelling out the least intuitive of these weak assumptions. The third weak assumption stipulated above is best illustrated by an example. To recapitulate, “typically” causal states are the triple  $(g_0, x_0, \tau_{0+})$ : hidden state, observed symbol, and time since last symbol. Usually,  $(g_0, x_0, \tau_{0+})$  and  $(g'_0, x'_0, \tau'_{0+})$  are not predictively equivalent when  $g_0 \neq g'_0$  and the uHSMM has the minimal number of states possible (the first weak assumption). However, to engineer such an equivalence without violating this first weak assumption, one first needs the two observed symbols to be identical,  $x_0 = x'_0$ . Next, one needs the hidden state to which we transition to be identical, so that:

$$p((x_1, \tau_1) | g_0, x_0, \tau_0) = p(x_1 | \epsilon(g_0, x_0)) \phi_{\epsilon(g_0, x_0)}(\tau_1).$$

And, finally, one needs:

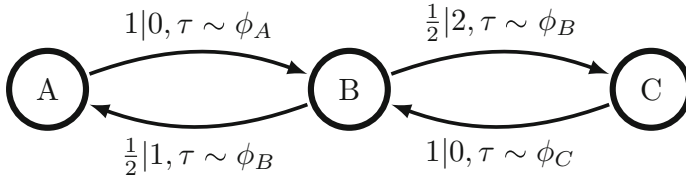
$$p(\tau_{0-} | g_0, \tau_{0+}) = \frac{\phi_{g_0}(\tau_{0+} + \tau_{0-})}{\Phi_{g_0}(\tau_{0+})}$$

to be equal to  $p(\tau_{0-} | g'_0, \tau'_{0+})$ , for all  $\tau_{0+} \geq 0$ . This last equality can be obtained in myriad ways. For instance, in the example of Fig. 2, we choose:

$$\phi_C(t) = \begin{cases} \left( \int_0^T \phi_A(t') dt' \right) \delta(t - \frac{T}{2}) & t \leq T \\ \phi_A(t) & t > T \end{cases} \tag{5}$$

Then, the triple  $(A, 0, t)$  is predictively equivalent to  $(C, 0, t)$  for  $t > T$  and the originally distinct conveyor belts corresponding to  $t$  “merge”.

Calculating information measures of a time series exactly often requires finding the probability distribution  $p(g_0, x_{0+}, \tau_{0+})$  over causal states. For instance, to find a process' statistical



**Fig. 2** With  $\phi_C(t)$  given by Eq. (5), the triple  $(A, 0, t)$  is predictively equivalent to  $(C, 0, t)$  for  $t > T$

complexity  $C_\mu$ , we must determine  $H[S]$  which, in turn, entails finding the probability distribution  $p(g_0, x_{0+}, \tau_{0+})$ . Implicitly, we are deriving labeled transition operators as in Ref. [25]. We start by decomposing:

$$p(g_0, x_{0+}, \tau_{0+}) = p(g_0)p(x_{0+}|g_0)p(\tau_{0+}|g_0, x_{0+}).$$

Since the dwell-time distribution depends only on the hidden state  $g_0$  and not on the emitted symbol  $x_{0+}$ , we find that:

$$p(\tau_{0+}|g_0, x_{0+}) = p(\tau_{0+}|g_0).$$

As in Ref. [25] in Sect. IVA, having a dwell time of at least  $\tau_{0+}$  implies that:

$$\begin{aligned} p(\tau_{0+}|g_0) &= \int_{\tau_{0+}}^{\infty} p(\tau_0|g_0)d\tau_0 \\ &= \mu_{g_0} \Phi_{g_0}(\tau_{0+}), \end{aligned} \tag{6}$$

where  $\Phi_{g_0}(\tau_{0+}) := \int_{\tau_{0+}}^{\infty} \phi_{g_0}(t)dt$  will be called the *survival distribution*, and:

$$\mu_{g_0} = 1 / \int_0^{\infty} t \phi_{g_0}(t)dt$$

is the inverse mean interevent interval from hidden state  $g_0$ . From the setup, we also have:

$$p(x_{0+}|g_0) = T_{g_0}^{(x_{0+})}.$$

Finally, to calculate  $p(g_0)$ , we consider all ways in which probability can flow from  $(g', x', \tau')$  to  $(g, x, 0)$ :

$$p(g, x, 0) = \sum_{g', x'} \int_0^{\infty} p(g', x', \tau') p((g', x', \tau') \rightarrow (g, x, 0)) d\tau'. \tag{7}$$

The transition probability  $p((g', x', \tau') \rightarrow (g, x, 0))$  is:

$$p((g', x', \tau') \rightarrow (g, x, 0)) = \delta_{g, \epsilon(g', x')} T_g^{(x)} \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')}. \tag{8}$$

The term  $\delta_{g, \epsilon(g', x')}$  implies that one can only transition to  $g$  from  $g'$  if the emitted symbol  $x'$  allows. Then,  $T_g^{(x)}$  implies that there is a probability of emitting symbol  $x$  from newly-transitioned-to hidden state  $g$ . And,  $\phi_{g'}(\tau') / \Phi_{g'}(\tau')$  is the probability of emitting  $x'$  for total time  $\tau'$ , given that  $x'$  has already been emitted for total time at least  $\tau'$ . Combining Eq. (8) with Eq. (7) gives:

$$\begin{aligned}
 p(g)T_g^{(x)}\mu_g &= \sum_{g',x'} \int_0^\infty p(g')T_{g'}^{(x')} \mu_{g'}\Phi_{g'}(\tau')\delta_{g,\epsilon(g',x')}T_g^{(x)} \frac{\phi_{g'}(\tau')}{\Phi_{g'}(\tau')} \\
 p(g) &= \frac{1}{\mu_g} \sum_{g',x'} \mu_{g'}p(g')T_{g'}^{(x')} \delta_{g,\epsilon(g',x')} \\
 &= \sum_{g'} \frac{\mu_{g'}}{\mu_g} \left( \sum_{x'} T_{g'}^{(x')} \delta_{g,\epsilon(g',x')} \right) p(g').
 \end{aligned}$$

We therefore see that  $p(g)$  is the eigenvector (appropriately normalized,  $\sum_g p(g) = 1$ ) associated with eigenvalue 1 of a transition matrix given by:

$$T_{g' \rightarrow g} := \frac{\mu_{g'}}{\mu_g} \left( \sum_{x'} T_{g'}^{(x')} \delta_{g,\epsilon(g',x')} \right). \tag{9}$$

Recalling computation theory [26], we see that the  $\epsilon$ -machine of such a process takes on the form of connected counters, albeit in continuous time. See Fig. 3.

A process' statistical complexity is defined as the entropy of its causal states. To calculate this, we need to find the probability distribution over the triple  $(g_0, x_{0+}, \tau_{0+})$ , from which we can compute the statistical complexity as stated in Proposition 1. In the proposition, we overloaded notation:  $H[p(g)] = -\sum_g p(g) \log p(g)$ .

**Proposition 1** *The statistical complexity of a unifilar hidden semi-Markov process, under weak assumptions, is given by:*

$$C_\mu = H[p(g)] - \sum_g p(g) \left( \sum_x T_g^{(x)} \log T_g^{(x)} \right) - \sum_g p(g) \int_0^\infty (\mu_g \Phi_g(\tau)) \log (\mu_g \Phi_g(\tau)) d\tau,$$

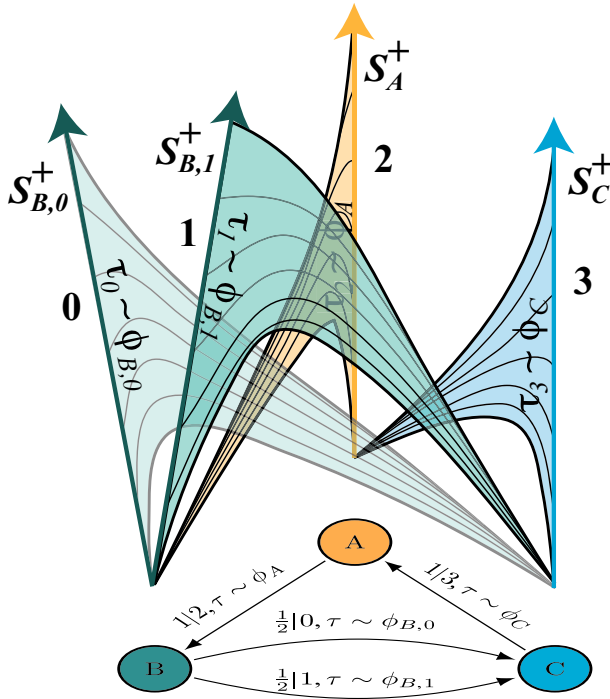
where, as above,  $p(g)$  is the normalized right eigenvector of eigenvalue 1 of the matrix of Eq. (9).

*Proof* Altogether, we find that the statistical complexity is:

$$\begin{aligned}
 C_\mu &= H[\mathcal{G}, X, T] \\
 &= H[\mathcal{G}] + H[X|\mathcal{G}] + H[T|\mathcal{G}, X] \\
 &= H[p(g)] - \sum_g p(g) \left( \sum_x T_g^{(x)} \log T_g^{(x)} \right) - \sum_g p(g) \int_0^\infty (\mu_g \Phi_g(\tau)) \log (\mu_g \Phi_g(\tau)) d\tau,
 \end{aligned}$$

where  $p(g)$  is given above.

As with continuous-time renewal processes, this is the statistical complexity of a mixed random variable and, hence, is not always an upper bound on the excess entropy  $\mathbf{E} := I \left[ \overleftarrow{(X, T)}; \overrightarrow{(X, T)} \right]$ .



**Fig. 3** Continuous-time  $\epsilon$ -machine for the hidden semi-Markov process generated by the uHSM of Fig. 1, as determined by Theorem 1. Continuous-time causal states  $S_A^+$ ,  $S_B^+$ , and  $S_C^+$  track the times  $\tau_0$  or  $\tau_1$  since last event—the times obeying distributions  $\phi_A$ ,  $\phi_{B,0}$ ,  $\phi_{B,1}$ , and  $\phi_C$  associated with uHMM states  $A$ ,  $B$ , and  $C$ , respectively. Each renewal subprocess is depicted as a semi-infinite vertical line and is isomorphic with the positive real line. If no event is seen, probability flows towards increasing time since last event, as described in Eq. (9). Otherwise, the surfaces leaving  $S_A^+$ ,  $S_B^+$ , and  $S_C^+$  indicate allowed transitions back to the next reset state or 0 node located at the nonarrow end of the associated  $S^+$ , denoting that a new event occurred associated with the next state  $A$ ,  $B$ , and  $C$ , as appropriate. Note that when leaving state  $B$  there are two distinct diffusion processes on  $S_B^+$  associated with emitting either 0 and 1. The domains of these diffusions are depicted with two separate semi-infinite lines, denoted  $S_{B,0}^+$  and  $S_{B,1}^+$ , respectively. Figure 1’s generative HSMM is displayed underneath for reference

### 4 Informational Architecture

Finally, we use the causal-state identification to calculate a uHSM’s entropy rate. Entropy rates are defined via:

$$h_\mu = \lim_{T \rightarrow \infty} \frac{H[(\overline{X}, \overline{T})^T]}{T},$$

but can be calculated using:

$$h_\mu = \lim_{\delta \rightarrow 0} \frac{H[(\overline{X}, \overline{T})^\delta | \overleftarrow{(\overline{X}, \overline{T})}]}{\delta}.$$

Starting there and recalling the definition of causal states, we immediately have:

$$h_\mu = \lim_{\delta \rightarrow 0} \frac{H[(\overline{X}, \overline{T})^\delta | \mathcal{S}]}{\delta}. \tag{10}$$



We find that entropy rate is as stated in Theorem 2. We could have just as well conditioned on prescient states [3], and so the weak assumptions of Theorem 1 are irrelevant to Theorem 2. The form of the expression appears as a weighted sum of differential entropies of dwell times from the hidden states.

**Theorem 2** *The entropy rate of a unifilar hidden semi-Markov process is given by:*

$$h_\mu = - \sum_g p(g) \int_0^\infty \mu_g \phi_g(\tau) \log \phi_g(\tau) d\tau,$$

where  $p(g)$  is the normalized right eigenvector associated with eigenvalue 1 of the matrix in Eq. (9).

*Proof* As just noted, the entropy rate can be calculated via:

$$h_\mu = \lim_{\delta \rightarrow 0} \frac{H[(\overrightarrow{X}, \overrightarrow{T})^\delta | \mathcal{S}]}{\delta},$$

where  $(\overrightarrow{X}, \overrightarrow{T})^\delta$  are trajectories of time length  $\delta$ . Following Ref. [25], we consider the random variable  $X_\delta$  to be 0 when the emitted symbol is constant throughout the trajectory of length  $\delta$ , 1 when there is one switch from one emitted symbol to another in this trajectory of length  $\delta$ , and so on. Basic information formulae give:

$$H[(\overrightarrow{X}, \overrightarrow{T})^\delta | \mathcal{S}] = H[X_\delta | \mathcal{S}] + H[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta, \mathcal{S}]. \tag{11}$$

The key reason that we condition on  $X_\delta$  is that two switches are highly unlikely to happen relative to one switch. Furthermore, if no switches in emitted symbols occur, then the trajectory is entirely predictable and does not contribute to the entropy rate. In particular:

$$\begin{aligned} \Pr(X_\delta = 1 | \mathcal{S} = (g, x, \tau)) &= \int_0^\delta \frac{\phi_g(\tau + s)}{\Phi_g(\tau + s)} ds \\ &\approx \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta + O(\delta^2) \quad \text{and} \\ \Pr(X_\delta = k | \mathcal{S} = (g, x, \tau)) &= O(\delta^k) \end{aligned}$$

and so:

$$\begin{aligned} \Pr(X_\delta = 0 | \mathcal{S} = (g, x, \tau)) &= 1 - \Pr(X_\delta \geq 1 | \mathcal{S} = (g, x, \tau)) \\ &= 1 - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta + O(\delta^2). \end{aligned}$$

In a straightforward way, it follows that:

$$\begin{aligned} H[X_\delta | \mathcal{S} = (g, x, \tau)] &= - \left( 1 - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \right) \log \left( 1 - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \right) \\ &\quad - \left( \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \right) \log \left( \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \right) + O(\delta^2 \log \delta) \\ &= \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta - \left( \frac{\phi_g(\tau)}{\Phi_g(\tau)} \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} \right) \delta - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \log \delta + O(\delta^2 \log \delta), \end{aligned} \tag{12}$$

after several Taylor approximations; e.g.,  $\log(1 + x) = x + O(x^2)$ .

Now, consider the second term in Eq. (11):

$$\begin{aligned} & \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta, \mathcal{S} = (g, x, \tau)] \\ &= \Pr(X_\delta = 0 | \mathcal{S} = (g, x, \tau)) \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = 0, \mathcal{S} = (g, x, \tau)] \\ & \quad + \Pr(X_\delta = 1 | \mathcal{S} = (g, x, \tau)) \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = 1, \mathcal{S} = (g, x, \tau)] \\ & \quad + \sum_{k=2}^\infty P(X_\delta = k | \mathcal{S} = (g, x, \tau)) \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = k, \mathcal{S} = (g, x, \tau)]. \end{aligned}$$

If  $X_\delta = 0$ , the trajectory is completely determined by  $\mathcal{S} = (g, x, \tau)$ , and hence:

$$\mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = 0, \mathcal{S} = (g, x, \tau)] = 0.$$

If  $X_\delta = 1$ , then the trajectory is completely determined by one time—that at which emitted symbols switch. As in Ref. [25], the distribution of switching time is roughly uniform over the interval, and so:

$$\mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = 1, \mathcal{S} = (g, x, \tau)] = \log \delta + O(\delta).$$

Finally, from maximum entropy arguments:

$$P(X_\delta = k | \mathcal{S} = (g, x, \tau)) \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = k, \mathcal{S} = (g, x, \tau)]$$

is at most of  $\delta^k (\log \delta)^k$ . In particular, we noted earlier that  $P(X_\delta = k | \mathcal{S} = (g, x, \tau))$  was  $O(\delta^k)$  and that  $k$  emissions over a time interval of no more than  $\delta$  yields differential entropy of no more than  $(\log \delta)^k$ . In addition, we have:

$$\mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = k] \geq \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = 1].$$

That is, if given a trajectory with a single transition, one can construct trajectories that approximate it arbitrarily closely with more than one transition. And so,  $|\mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = k]|$  is at least  $O(|\log \delta|)$  and at most  $O(|\log \delta|^k)$ . Hence:

$$\sum_{k=2}^\infty P(X_\delta = k | \mathcal{S} = (g, x, \tau)) \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | X_\delta = k, \mathcal{S} = (g, x, \tau)] = O(\delta^2 (\log \delta)^2).$$

Altogether, we have:

$$\begin{aligned} \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | \mathcal{S} = (g, x, \tau)] &= \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta - \left( \frac{\phi_g(\tau)}{\Phi_g(\tau)} \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} \right) \delta - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \log \delta \\ & \quad + \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta \log \delta + O(\delta^2 (\log \delta)^2) \\ &= \frac{\phi_g(\tau)}{\Phi_g(\tau)} \delta - \left( \frac{\phi_g(\tau)}{\Phi_g(\tau)} \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} \right) \delta + O(\delta^2 (\log \delta)^2) \end{aligned}$$

and, thus:

$$\begin{aligned} \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | \mathcal{S}] &= \langle \mathbb{H}[(\overrightarrow{X}, \overrightarrow{T})^\delta | \mathcal{S} = (g, x, \tau)] \rangle_{g,x,\tau} \\ &= \left\langle \frac{\phi_g(\tau)}{\Phi_g(\tau)} - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} \right\rangle_{g,x,\tau} \delta + O(\delta^2 (\log \delta)^2). \end{aligned}$$

And so, from Eq. (10), we find the entropy rate:

$$\begin{aligned}
 h_\mu &= \lim_{\delta \rightarrow \infty} \frac{H[(\overline{X, T})^\delta | \mathcal{S}]}{\delta} \\
 &= \left\langle \frac{\phi_g(\tau)}{\Phi_g(\tau)} - \frac{\phi_g(\tau)}{\Phi_g(\tau)} \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} \right\rangle_{g,x,\tau} \\
 &= \sum_{g,x} p(g) T_g^{(x)} \int_0^\infty \mu_g \Phi_g(\tau) \frac{\phi_g(\tau)}{\Phi_g(\tau)} d\tau - \sum_{g,x} p(g) T_g^{(x)} \int_0^\infty \mu_g \Phi_g(\tau) \frac{\phi_g(\tau)}{\Phi_g(\tau)} \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} \\
 &= \sum_{g,x} p(g) T_g^{(x)} \left( \int_0^\infty \mu_g \phi_g(\tau) d\tau - \int_0^\infty \mu_g \phi_g(\tau) \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} d\tau \right).
 \end{aligned}$$

We directly have:

$$\int_0^\infty \mu_g \phi_g(\tau) d\tau = \mu_g$$

and

$$\int_0^\infty \mu_g \phi_g(\tau) \log \frac{\phi_g(\tau)}{\Phi_g(\tau)} = \int_0^\infty \mu_g \phi_g(\tau) \log \phi_g(\tau) d\tau - \int_0^\infty \mu_g \phi_g(\tau) \log \Phi_g(\tau) d\tau.$$

The second term simplifies substituting  $u = \Phi_g(\tau)$ :

$$\begin{aligned}
 \int_0^\infty \mu_g \phi_g(\tau) \log \Phi_g(\tau) d\tau &= - \int_1^0 \mu_g \log u du \\
 &= \mu_g (u \log u - u) \Big|_0^1 \\
 &= -\mu_g.
 \end{aligned}$$

Altogether, we find:

$$h_\mu = - \sum_g p(g) \int_0^\infty \mu_g \phi_g(\tau) \log \phi_g(\tau) d\tau.$$

Theorem 2 generalizes a recent result about the entropy rate of semi-Markov processes [22] to unifilar hidden semi-Markov processes. In this way, Theorem 2 demonstrates the practical use of identifying causal states.

### 5 Conclusion

Proposition 1 and Theorem 2 provide new expressions for the statistical complexity and entropy rate of continuous-time, discrete-event processes generated by a restricted class of hidden semi-Markov models, what we called “unifilar hidden semi-Markov models”. Their causal architecture reminds one of a discrete-time  $\epsilon$ -machine crossed with the  $\epsilon$ -machine of a continuous-time renewal process [25]. Pursuing this, straightforward extensions of the techniques in Ref. [25] allowed us to derive new expressions for statistical complexity and entropy rate.

These expressions can be used as new plug-in estimators for the statistical complexity and entropy rate of continuous-time, discrete-event processes. It might seem that using them requires an accurate estimation of  $\phi_g(\tau)$ , which in turn might require unreasonable amounts of data. However, a method first utilized by Kozachenko and Leonenko and popularized by

Victor [27] utilizes the fact that we need only calculate scalar functions of  $\phi_g(\tau)$  and not  $\phi_g(\tau)$  itself. A second concern arises from the super-exponential explosion of discrete-time, discrete-alphabet  $\epsilon$ -machines with the number of hidden states [28]. How do we know the underlying topology? Here, as a first try, we suggest taking Ref. [29]'s approach, replacing the hidden states in these general models with the last  $k$  symbols. Further research is required, though, to determine when the chosen  $k$  is too small or too large. And, finally, we hope this work stimulates new work on continuous-time, discrete-alphabet quantum machines such as found in Ref. [30].

**Acknowledgements** The authors thank Santa Fe Institute for its hospitality during visits, A. Boyd, C. Hillar, and D. Upper for useful discussions, and T. Elliott for the example of Fig. 2. JPC is an SFI External Faculty member. This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U. S. Army Research Office under Contracts W911NF-13-1-0390 and W911NF-12-1-0288. S.E.M. was funded by a National Science Foundation Graduate Student Research Fellowship, a U.C. Berkeley Chancellor's Fellowship, and the MIT Physics of Living Systems Fellowship.

## References

1. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(379–423), 623–656 (1948)
2. Crutchfield, J.P., Young, K.: Inferring statistical complexity. *Phys. Rev. Lett.* **63**, 105–108 (1989)
3. Shalizi, C.R., Crutchfield, J.P.: Computational mechanics: pattern and prediction, structure and simplicity. *J. Stat. Phys.* **104**, 817–879 (2001)
4. Varn, D.P., Crutchfield, J.P.: Chaotic crystallography: how the physics of information reveals structural order in materials. *Curr. Opin. Chem. Eng.* **7**, 47–56 (2015)
5. Kelly, D., Dillingham, M., Hudson, A., Wiesner, K.: A new method for inferring hidden Markov models from noisy time sequences. *PLoS ONE* **7**(1), e29703 (2012)
6. Li, C.-B., Yang, H., Komatsuzaki, T.: Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl. Acad. Sci. USA* **105**, 536–541 (2008)
7. Li, C.-B., Komatsuzaki, T.: Aggregated Markov model using time series of a single molecule dwell times with a minimum of excessive information. *Phys. Rev. Lett.* **111**, 058301 (2013)
8. Marzen, S., DeWeese, M.R., Crutchfield, J.P.: Time resolution dependence of information measures for spiking neurons: scaling and universality. *Front. Comput. Neurosci.* **9**, 109 (2015)
9. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. *Nature* **453**(7196), 779–782 (2008)
10. Witt, A., Neiman, A., Kurths, J.: Characterizing the dynamics of stochastic bistable systems by measures of complexity. *Phys. Rev. E* **55**, 5050–5059 (1997)
11. Clarke, R.W., Freeman, M.P., Watkins, N.W.: Application of computational mechanics to the analysis of natural data: an example in geomagnetism. *Phys. Rev. E* **67**, 016203 (2003)
12. Dzugutov, M., Aurell, E., Vulpiani, A.: Universal relation between the Kolmogorov-Sinai entropy and the thermodynamical entropy in simple liquids. *Phys. Rev. Lett.* **81**, 1762 (1998)
13. Gonçalves, W.M., Pinto, R.D., Sartorelli, J.C., de Oliveira, M.J.: Inferring statistical complexity in the dripping faucet experiment. *Physica A* **257**(1–4), 385–389 (1998)
14. Jay Palmer, A., Fairall, C.W., Brewer, W.A.: Complexity in the atmosphere. *IEEE Trans. Geosci. Remote Sens.* **38**, 2056–2063 (2000)
15. Cerbus, R.T., Goldberg, W.I.: Information content of turbulence. *Phys. Rev. E* **88**, 053012 (2013)
16. Kolmogorov, A.N.: *Foundations of the Theory of Probability*, 2nd edn. Chelsea Publishing Company, New York (1956)
17. Chaitin, G.: On the length of programs for computing finite binary sequences. *J. ACM* **13**, 145 (1966)
18. Crutchfield, J.P.: Between order and chaos. *Nat. Phys.* **8**(January), 17–24 (2012)
19. Akimoto, T., Hasumi, T., Aizawa, Y.: Characterization of intermittency in renewal processes: application to earthquakes. *Phys. Rev. E* **81**, 031133 (2010)
20. Darmon, D., Sylvestre, J., Girvan, M., Rand, W.: Predictability of user behavior in social media: bottom-up versus top-down modeling. [arXiv:1306.6111](https://arxiv.org/abs/1306.6111)
21. Yu, S.-Z.: Hidden semi-Markov models. *Artif. Intell.* **174**(2), 215–243 (2010)

22. Girardin, V.: On the different extensions of the ergodic theorem of information theory. In: Baeza-Yates, R., Glaz, J., Gzyl, H., Husler, J., Palacios, J.L. (eds.) *Recent Advances in Applied Probability Theory*, pp. 163–179. Springer, New York (2005)
23. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*, 2nd edn. Wiley, New York (2006)
24. Nair, C., Prabhakar, B., Shah, D.: On entropy for mixtures of discrete and continuous variables (2006). [arXiv:cs/0607075](https://arxiv.org/abs/cs/0607075)
25. Marzen, S., Crutchfield, J.P.: Informational and causal architecture of continuous-time renewal processes. *J. Stat. Phys.* **168**(1), 109–127 (2017)
26. Brookshear, J.G.: *Theory of Computation: Formal Languages, Automata, and Complexity*. Benjamin/Cummings, Redwood City (1989)
27. Victor, J.D.: Binless strategies for estimation of information from neural data. *Phys. Rev. E* **66**(5), 051903 (2002)
28. Johnson, B.D., Crutchfield, J.P., Ellison, C.J., McTague, C.S.: Enumerating finitary processes. [arXiv:1011.0036](https://arxiv.org/abs/1011.0036)
29. Strelhoff, C.C., Crutchfield, J.P., Hübler, Alfred: Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Phys. Rev. E* **76**(1), 011106 (2007)
30. Elliot, T.J., Gu, M.: Occam’s Vorpall Quantum Razor: Memory Reduction when Simulating Continuous-time Stochastic Processes with Quantum Devices (2017). [arXiv:1704.04231](https://arxiv.org/abs/1704.04231)