Characterizing information-theoretic storage and transfer in continuous time processes

Richard E. Spinney and Joseph T. Lizier
Complex Systems Research Group and Centre for Complex Systems, Faculty of Engineering and Information Technologies, University of Sydney, Sydney, New South Wales 2006, Australia

(Received 22 April 2018; published 23 July 2018)

The characterization of information processing is an important task in complex systems science. *Information dynamics* is a quantitative methodology for modeling the intrinsic information processing conducted by a process represented as a time series, but to date has only been formulated in discrete time. Building on previous work which demonstrated how to formulate transfer entropy in continuous time, we give a total account of information processing in this setting, incorporating information storage. We find that a convergent rate of predictive capacity, comprising the transfer entropy and active information storage, does not exist, arising through divergent rates of active information storage. We identify that active information storage can be decomposed into two separate quantities that characterize predictive capacity stored in a process: active memory utilization and instantaneous predictive capacity. The latter involves prediction related to path regularity and so solely inherits the divergent properties of the active information storage, while the former permits definitions of pathwise and rate quantities. We formulate measures of memory utilization for jump and neural spiking processes and illustrate measures of information processing in synthetic neural spiking models and coupled Ornstein-Uhlenbeck models. The application to synthetic neural spiking models demonstrates that active memory utilization for point processes consists of discontinuous jump contributions (at spikes) interrupting a continuously varying contribution (relating to waiting times between spikes), complementing the behavior previously demonstrated for transfer entropy in these processes.

DOI: 10.1103/PhysRevE.98.012314

I. INTRODUCTION

Information dynamics [1–4] is a framework that seeks to characterize distributed computation by identifying primitives of information processing in autonomously evolving or "computing" systems. Specifically, this involves modeling how information is stored in and transferred between variables in the system when the values of these variables are dynamically updated in time. These primitive information processing operations studied by the framework, information storage and transfer, are measured by the active information storage [4] and transfer entropy [5] (and higher orders thereof in larger multivariate systems [3,6]), respectively. The measures in this framework have been used to characterize and explain behavior observed in various complex systems, providing novel insights across a broad range of fields including canonical complex systems such as cellular automata [6] and random Boolean networks [7], interpretation of dynamics in [8] and improved algorithms for machine learning [9], characterizing information processing signatures in biological signaling networks [10], collective behavior in swarms [11,12], nonlinear time series forecasting [13], and computational neuroscience applications identifying neural information flows from brain imaging data [14–16], inferring effective network structure [17–19], providing evidence for the predictive coding hypothesis [20], and identifying differences in predictive information in autism spectrum disorder subjects and controls [21].

However, to date, no complete theoretical account for how this framework should be applied to continuous time systems has been offered, despite such systems being ubiquitous in fields throughout all of science. Recently we have given an account of how to formulate transfer entropy in such systems [22]. In this paper we build upon these developments by addressing the concept of predictive capacity, comprising the transfer entropy and active information storage, in these systems. We find that such a quantity is much more complicated owing to the predictive properties that can be derived from the regularity properties of continuous time processes. As such, we focus in this paper on investigating the nature of the active information storage in continuous time processes. To proceed we find that it is necessary to decompose the active information storage into two components related to two distinct characteristics of information processing. We call these two quantities *active memory utilization* and *instantaneous predictive capacity*. The former is designed to behave as a rate and to be complementary to transfer entropy. Meanwhile the latter relates intrinsic uncertainty and path regularity properties of the process. This may be interpreted asymptotically and independently of both the transfer entropy and active memory utilization. The results bring our understanding of the nature of information storage in such systems in line with that of information transfer, which is important not only in that they are both fundamental components of models of information processing, but also because of the insight these results bring to the growing role of information storage analyses in neural imaging data in particular [14,20,21]. Our insights reveal, for example, the expected behavior of active information storage and its components when measured on discrete time samples of underlying continuous time processes, whereby the active memory utilization is the only quantity that will approach a limiting value as the discrete time step approaches zero; this has significant implications for empirical analyses.

The overall plan of the paper is as follows. In Sec. II we introduce information dynamics as currently conceptualized. Following on from this in Sec. III we discuss information-theoretic measures in continuous time and outline the central problem under consideration, the definition and identification of predictive information associated with information storage in continuous time. Next in Sec. IV we present a set of postulates from which we deduce the central division of the active information storage, identifying the active memory utilization and instantaneous predictive capacity before describing their general behavior. In Sec. V we then detail the relation between the these quantities and other information-theoretic measures of stochastic processes such as excess entropy and differential entropy measures on continuous state spaces. Finally, before concluding in Sec. VIII, we present applications and examples chosen to complement previous demonstrations of the properties of the transfer entropy in continuous time [22,23]. In Sec. VI such quantities are presented for neural spiking models including analytical results for some simple synthetic models followed by a numerical full information processing description of a more complicated model comprising two spike trains, consisting of both active memory utilization and transfer entropy. In Sec. VII a coupled Ornstein-Uhlenbeck process is discussed where analytical results for transfer entropy and active memory utilization can be derived.

II. INFORMATION DYNAMICS

First, we summarize the basic principles of information dynamics as formulated in discrete time. Central to the concept of information dynamics is the idea that each evolving state can be modeled as being "computed" from the past, in the sense of an "intrinsic computation" [24], and that this computation is characterized by the *predictive capacity*. This concept is made concrete in the context of two random processes $X_{\{0:m\}} = \{X_0, \ldots, X_m\}$ and $Y_{\{0:m\}} = \{Y_0, \ldots, Y_m\}$ taking individual outcomes $x_{\{0:m\}} = \{x_0, \ldots, x_m\}$ and $y_{\{0:m\}} = \{y_0, \ldots, y_m\}$, wherein the predictive capacity of the state X_{n+1} at time $n < m$ is considered, axiomatically, to be the reduction of uncertainty in X_{n+1} that arises from knowing the path histories $X_{\{0:n\}} = x_{\{0:n\}}$ and $Y_{\{0:n\}} = y_{\{0:n\}}$, at time n, over having no other knowledge.

This predictive capacity is then formalized mathematically as C_X, given as a mutual information or the difference between two (conditional) entropies [3,6]

$$\begin{aligned} C_X &\equiv I(X_{n+1}; X_{\{0:n\}}, Y_{\{0:n\}}) \\ &= H(X_{n+1}) - H(X_{n+1}|X_{\{0:n\}}, Y_{\{0:n\}}) \end{aligned} \tag{1}$$

based on underlying ensemble probabilities, p. Following on from this quantification of predictive capacity, the central approach and framework of information dynamics is to decompose such a quantity into two specific terms with this decomposition termed the *computational signature*, viz.,

$$\begin{aligned} C_X &= H(X_{n+1}) - H(X_{n+1}|X_{\{0:n\}}) \\ &\quad + H(X_{n+1}|X_{\{0:n\}}) - H(X_{n+1}|X_{\{0:n\}}, Y_{\{0:n\}}) \\ &= I(X_{n+1}; X_{\{0:n\}}) + I(X_{n+1}; Y_{\{0:n\}}|X_{\{0:n\}}) \\ &= A_X + T_{Y\to X}, \end{aligned} \tag{2}$$

where A_X is known as the *active information storage* [4] and is explicitly written

$$\begin{aligned} A_X &\equiv H(X_{n+1}) - H(X_{n+1}|X_{\{0:n\}}) \\ &= \mathbb{E}\left[\ln \frac{p(x_{n+1}|x_{\{0:n\}})}{p(x_{n+1})}\right], \end{aligned} \tag{3}$$

and where $T_{Y\to X}$ is well known in many distinct areas of science as the *transfer entropy* [5,25–27] and is explicitly written

$$\begin{aligned} T_{Y\to X} &\equiv H(X_{n+1}|X_{\{0:n\}}) - H(X_{n+1}|X_{\{0:n\}}, Y_{\{0:n\}}) \\ &= \mathbb{E}\left[\ln \frac{p(x_{n+1}|x_{\{0:n\}}, y_{\{0:n\}})}{p(x_{n+1}|x_{\{0:n\}})}\right]. \end{aligned} \tag{4}$$

Here we note the shorthand $p(x_{n+1}|x_{\{0:n\}}, y_{\{0:n\}}) = p(X_{n+1} = x_{n+1}|X_{\{0:n\}} = x_{\{0:n\}}, Y_{\{0:n\}} = y_{\{0:n\}})$, etc., which we use to indicate, where appropriate, that the arguments of the probabilities are simply specific realizations of the variables to which the probabilities correspond, and the notation $\mathbb{E}[\ldots]$ which denotes an ensemble average with respect to $p(x_{\{0:n+1\}}, y_{\{0:n\}})$. Since both contributions can be written as (conditional) mutual informations both are rigorously non-negative.

The computational signature effects a model of intrinsic computation based on this partition into A_X and $T_{Y\to X}$, such that there exists an identifiable storage component attributed to the past of X through A_X, plus an information transfer component attributed to the past of Y in the context of X through $T_{Y\to X}$ [6].

More recent developments have emphasized that information-theoretic quantities should be constructed from suitable *local* or *pointwise* quantities, of which the ensemble quantities, A_X, $T_{Y\to X}$, etc., are suitable expectations. Consequently we recognize the structure of the *local active information storage*, a_X, given by $A_X = \mathbb{E}[a_X(x_{n+1}, x_{\{0:n\}})]$ [4] and *local transfer entropy*, $t_{Y\to X}$, given by $T_{Y\to X} = \mathbb{E}[t_{Y\to X}(x_{n+1}, x_{\{0:n\}}, y_{\{0:n\}})]$ [1]. Explicitly, we have

$$\begin{aligned} a_X(x_{n+1}, x_{\{0:n\}}) &\equiv \ln \frac{p(x_{n+1}|x_{\{0:n\}})}{p(x_{n+1})}, \\ t_{Y\to X}(x_{n+1}, x_{\{0:n\}}, y_{\{0:n\}}) &\equiv \ln \frac{p(x_{n+1}|x_{\{0:n\}}, y_{\{0:n\}})}{p(x_{n+1}|x_{\{0:n\}})}. \end{aligned} \tag{5}$$

We then also define the total computational signature, on a local scale, $c_X = a_X + t_{Y\to X}$. It is important to note that such local values have no bound on their sign and thus may be negative. Such an approach allows significance to placed on *single realizations* of a process, allowing fine characterization of spatial temporal features, such as the identification of dynamics that are informative, but especially those which are *misinformative*, characterized by negative local values (which have been shown to identify interesting aspects of dynamics in cellular automata [1,4] and stimulus changes in the cat visual cortex [28]).

III. FORMULATING QUANTITIES IN CONTINUOUS TIME

A. Background and established quantities

As originally formulated, information dynamics treats only systems in discrete time. However, many systems of great interest, not only to complexity research but to many areas

of science, are naturally formulated in continuous time. In such systems time series are not indexed by the integers, but by connected subsets of the real line such that instead of collections of random variables, one deals with random functions for which we use notation $X_{\mathcal{A}} = \{X(t') : t' \in \mathcal{A}\}$ with individual realizations $x_{\mathcal{A}}$. In previous work [22] we established how to treat the transfer entropy in continuous time. Important consequences of this work were the recognition that one must consider a transfer entropy *rate*, and that this is formulated from an expectation of a *functional* that assigns a number to individual realizations of the process called the *pathwise transfer entropy*, $\mathcal{T}_{Y\to X}^{[t_0,t]}[x_{[\tau,t]}, y_{[\tau,t)}]$, which represents the total accumulated predictive capacity transferred from Y in the context of the history of X, on the interval $[t_0,t]$ as a function of the path realizations $x_{[\tau,t]}$ and $y_{[\tau,t)}$, with $\tau \leqslant t_0 < t$. Importantly, these are constructed from probability measures on *complete paths*, emphasizing that quantifications of evolving sequences should make sense quantitatively and conceptually in the wider context of them being understood, rigorously, as outcomes of fully realized stochastic processes. We may summarize this through the expressions for the transfer entropy rate and its relation to the pathwise transfer entropy, for a process with a defined time origin at τ,

$$\begin{aligned} T_{Y\to X}^{[t_0,t]} &\equiv \mathbb{E}\left[\mathcal{T}_{Y\to X}^{[t_0,t]}[x_{[\tau,t]}, y_{[\tau,t)}]\right] \\ &= \int_{t_0}^{t} \dot{T}_{Y\to X}(t')dt' \\ \dot{T}_{Y\to X}(t) &\equiv \frac{d}{dt}\mathbb{E}\left[\mathcal{T}_{Y\to X}^{[t_0,t]}[x_{[\tau,t]}, y_{[\tau,t)}]\right] \\ \mathcal{T}_{Y\to X}^{[t_0,t]}[x_{[\tau,t]}, y_{[\tau,t)}] &\equiv \ln \frac{d\mathbb{P}_{X|\{Y\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}, \{y_{[\tau,t)}\}\right]}{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}, \end{aligned} \tag{6}$$

where $T_{Y\to X}^{[t_0,t]}$ is the expected, cumulative, "transferred" information on the interval $[t_0,t]$ equal to the expected pathwise transfer entropy or the integral of the transfer entropy rate, $\dot{T}_{Y\to X}(t)$, over the interval. Again, $\mathbb{E}$ indicates an ensemble expectation. This property arises from the pathwise transfer entropy explicitly being the logarithm of a Radon-Nikodym (RN) derivative between probability measures, $\mathbb{P}$, on $x_{(t_0,t]}$. The central quantity, the pathwise transfer entropy, exists when such an RN derivative exists and the measures are absolutely continuous with respect to each other. One may, nonrigorously but safely in most instances, consider such a quantity to be the ratio of path "probabilities" defined through

$$\begin{aligned} &\frac{d\mathbb{P}_{X|\{Y\}}[x_{(t_0,t]}|x_{[\tau,t_0]}, \{y_{[\tau,t)}\}]}{d\mathbb{P}_X[x_{(t_0,t]}|x_{[\tau,t_0]}]} \\ &\quad\sim \lim_{n\to\infty} \prod_{i=0}^{n} \frac{p(x_{i+1}|x_{\{-k:i\}}, y_{\{-k:i\}})}{p(x_{i+1}|x_{\{-k:i\}})}, \end{aligned} \tag{7}$$

where $t_0 = 0$, $x_i \equiv x_{i\Delta t}$, and $\Delta t = t/(n+1) = -\tau/k$. We note the construction of $\mathbb{P}_{X|\{Y\}}[x_{(t_0,t]}|x_{[\tau,t_0]}, \{y_{[\tau,t)}\}] \sim \prod_{i=0}^{n} p(x_{i+1}|x_{\{-k:i\}}, y_{\{-k:i\}})$, emphasizing that this asymptotic product form in general is not equal to the analogously constructed usual conditional probability $p(x_{\{1:n+1\}}|x_{\{-k:0\}}, y_{\{-k:n\}})$.

When dealing with integrated quantities such as the pathwise transfer entropy, which represents the accumulated transfer of information on the interval, one may construct a rate in an alternative sense, viz.,

$$\mathring{T}_{Y\to X} \equiv \lim_{t\to\infty} \frac{1}{t-t_0}\mathbb{E}\left[\ln \frac{d\mathbb{P}_{X|\{Y\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}, \{y_{[\tau,t]}\}\right]}{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}\right]. \tag{8}$$

When the process is stationary we have $\mathring{T}_{Y\to X} = \dot{T}_{Y\to X}$. Moreover, in addition, when the process is ergodic it may be expressed

$$\mathring{T}_{Y\to X} = \lim_{t\to\infty} \frac{1}{t-t_0} \ln \frac{d\mathbb{P}_{X|\{Y\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}, \{y_{\tau,t)}\}\right]}{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}, \tag{9}$$

forming the basis for any empirical estimation methods.

B. Extension to active information storage and issues

The primary motivation of this paper is to extend the full description of information processing in intrinsic computation, outlined in the previous section, to the case of continuous time. To do so in keeping with the above definitions of the transfer entropy rate is straightforward: the predictive capacity and subsequent computational signature is defined over some time interval Δt, with mean rates emerging when dividing by Δt as the limit $\Delta t \to 0$ is taken. First, the local predictive capacity, before the notion of a rate introduced, is therefore captured by

$$\begin{aligned} c_X &\equiv \lim_{\Delta t\to 0} \ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]}, y_{[\tau,t]})}{p(x_{t+dt})} \\ &= \lim_{\Delta t\to 0} \ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t})} \\ &\quad + \ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]}, y_{[\tau,t]})}{p(x_{t+\Delta t}|x_{[\tau,t]})} \\ &= a_X + t_{Y\to X}. \end{aligned} \tag{10}$$

A *rate* of (average) predictive capacity would be then constructed as follows:

$$\begin{aligned} \dot{C}_X &\equiv \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]}, y_{[\tau,t]})}{p(x_{t+\Delta t})}\right] \\ &= \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t})}\right] \\ &\quad + \frac{1}{\Delta t}\mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]}, y_{[\tau,t]})}{p(x_{t+\Delta t}|x_{[\tau,t]})}\right] \\ &= \dot{A}_X + \dot{T}_{Y\to X}, \end{aligned} \tag{11}$$

comprising the previously defined transfer entropy rate $\dot{T}_{Y\to X}$ and an analogously defined "active information storage rate" $\dot{A}_X$. However, there are significant issues associated with this quantity, $\dot{C}_X$, stemming from the proposed contribution, $\dot{A}_X$.

To understand the problem with such a quantity it is helpful to reconstruct it in the manner of the definition of the transfer entropy rate in Eq. (6) through the definition of a hypothetical "pathwise active information storage," $\mathcal{A}_X^{[t_0,t]}[x_{[\tau,t]}]$, viz.,

$$\mathcal{A}_X^{[t_0,t]}[x_{[\tau,t]}] \equiv \ln \frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^{\emptyset}\left[x_{(t_0,t]}\right]} \tag{12}$$

with

$$\frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^{\emptyset}\left[x_{(t_0,t]}\right]} \sim \lim_{n\to\infty}\prod_{i=0}^{n}\frac{p(x_{i+1}|x_{\{-k:i\}})}{p(x_{i+1})}. \tag{13}$$

Here the following issue emerges: this quantity does not converge since the limit of the denominator does not lead to a measure on $x_{(t_0,t]}$; i.e., $\mathbb{P}_X^{\emptyset}$ does not exist. It then follows that a finite active information storage rate, $\dot{A}_X$, does not exist either. It is instructive to demonstrate this, and the general problem, with a brief example.

To demonstrate the issue with identifying predictive capacities and active information storage as currently defined in continuous time we consider a simple Ornstein-Uhlenbeck process in stochastic differential equation form,

$$dx_t = -\kappa x_t dt + \sigma dW_t, \tag{14}$$

where W_t is a Wiener process. Since there is no extrinsic process (i.e., Y) to consider, the total predictive capacity of the intrinsic computation is identical to the active information storage. When formulating such a prediction we may leave the time horizon over which such a prediction is made to be a free parameter, Δt. Doing so leads to a parametric predictive capacity or active information storage, $C_X^{(\Delta t)} = A_X^{(\Delta t)}$, given by

$$\begin{aligned} A_X^{(\Delta t)} &\equiv \lim_{(t-\tau)\to\infty}\mathbb{E}\left[\ln\frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t})}\right] \\ &= \mathbb{E}\left[\ln\frac{p(x_{t+\Delta t}|x_t)}{p(x_{t+\Delta t})}\right] \\ &= \frac{1}{2}\ln\left[\frac{e^{\kappa\Delta t}}{e^{\kappa\Delta t}-e^{-\kappa\Delta t}}\right], \end{aligned} \tag{15}$$

which is easily calculated using the well known transition probability density of the Ornstein-Uhlenbeck process [29], detailed in Appendix A 1, where it has been assumed that the process is in the stationary state. If we attempt to construct a rate in the limit $\Delta t \to 0$, it diverges. Moreover, even an attempt to find a convergent $O(1)$ quantity in the limit $\Delta t \to 0$ fails. This has a simple interpretation: with knowledge of the process's history, the sampling paths of the process allow for arbitrary precision in the prediction at smaller and smaller time horizons. This is reflected in the numerator of a_X which tends to a delta function. However, this contribution is simply unmatched by the uncertainty without conditioning on the process's history appearing in the denominator, which, as a Shannon (differential) entropy, remains bounded independently of the prediction horizon.

In this sense we see that the active information storage is actually performing precisely as it should: one *can* be infinitely more precise over an infinitesimal time horizon with knowledge of the process's history than with an isolated prediction for such processes. This motivates many additional questions: is such a measure of stored predictive information appropriate? Can it be decomposed into divergent and nondivergent terms in the $\Delta t \to 0$ limit and do these quantities possess sensible interpretations that can be meaningfully explored? We take the position that the active information storage decomposes into two distinct quantities related to active memory utilization and instantaneous predictive capacity. These are detailed in the subsequent sections.

IV. DECOMPOSITION OF STORED INFORMATION INTO ACTIVE MEMORY UTILIZATION AND INSTANTANEOUS PREDICTIVE CAPACITY

Here we posit that in general the active information storage A_X describes a generalized sense of memory utilization and is, in fact, composed of two quantities. The first is related to memory, understood in an intuitive manner, while the second does not characterize memory, but the predictive capacity that is obtained solely from the current state of the system we term "instantaneous prediction." We will demonstrate that these, in general, describe two distinct features of stochastic processes. The quantity that describes memory is a dynamical quantity that possesses a rate which we call the *active memory utilization rate*, $\dot{M}_X$. The instantaneous prediction is a nondynamical quantity not amenable to description as a rate which we call the *instantaneous predictive capacity*, I_X, where

$$A_X = I_X + \dot{M}_X\Delta t + O(\Delta t^2), \tag{16}$$

with $I_X \geqslant 0$ and $\dot{M}_X \geqslant 0$. We point out that $\dot{M}dt$ need not, however, comprise all $O(dt)$ contributions in A_X, since I_X may have $O(dt)$ components also. We arrive at such a division through the introduction of the following postulates designed to construct a quantity $\dot{M}_X$ which is complementary to $\dot{T}_{Y\to X}$:

(1) Measures of memory utilization should assign finite, unitless values to complete path realizations of a time series or stochastic process.

(2) Measures of memory utilization should be formed from RN derivatives between equivalent probability measures[1] on path realizations.

(3) The active information storage *contains* memory utilization such that any decomposition yields positive quantities in expectation.

(4) The memory utilization is found by *maximizing* such a component of active information storage such that the first two postulates are met.

Informally, this is to be interpreted as the requirement that, regardless of time basis, we should (i) be able to discuss the memory that has been cumulatively utilized over finite intervals, (ii) measure memory by comparing the relative weight assigned to the paths over these intervals by two models of the behavior which agree on which paths are possible, and (iii) find the largest mean contribution that achieves this with the currently existing measure of information storage which is derived from the (axiomatically fundamental) predictive capacity.

In order to meet the second and third postulates, we must decompose A_X through the introduction of a new transition probability, $\mathcal{P}$, that converges to a measure which is equivalent to $p[x_{[t_0,t]}]$, but also describes the statistics of the process such that it is an ensemble probability, p, itself. This ensures that it is a component of A_X, ensuring both $I_X \geqslant 0$ and $\dot{M}_X \geqslant 0$. To achieve the first two postulates we must retain the path regularity of the process and thus include x_t in the condition of

[1]Equivalent probability measures are those which are absolutely continuous with respect to each other such that they agree on which sets of events have zero probability. Informally, this is to be understood as them agreeing on which realizations are "possible."

the transition probability, along with an unspecified additional component $\mathcal{A} \subseteq x_{[\tau,t)}$ required for positivity related to the third postulate. To achieve this we write

$$\begin{aligned} A_X &= \lim_{\Delta t\to 0} \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t})}\right] \\ &= \lim_{\Delta t\to 0} \mathbb{E}\left[\ln \frac{\mathcal{P}(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t})}\right] + \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{\mathcal{P}(x_{t+\Delta t}|x_{[\tau,t]})}\right] \\ &= \lim_{\Delta t\to 0} \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_t \cup \{\mathcal{A} \subseteq x_{[\tau,t)}\})}{p(x_{t+\Delta t})}\right] \\ &\quad + \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t}|x_t \cup \{\mathcal{A} \subseteq x_{[\tau,t)}\})}\right] \\ &= I_X + \dot{M}_X \Delta t + O(\Delta t^2). \end{aligned} \tag{17}$$

The portion explainable as a dynamic memory source, $\dot{M}_X$, is then separated from the remainder I_X by maximizing $\dot{M}_X$ with respect to $\mathcal{A}$ such that we have

$$\begin{aligned} \dot{M}_X &\equiv \max_{\mathcal{A}} \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t}|x_t \cup \{\mathcal{A} \subseteq x_{[\tau,t)}\})}\right] \\ &= \lim_{\Delta t\to 0} \frac{1}{\Delta t}\mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t}|x_t)}\right] \end{aligned} \tag{18}$$

corresponding to $\mathcal{A} = \emptyset$ and $\mathcal{P}(x_{t+\Delta t}|x_{[\tau,t]}) = p(x_{t+\Delta t}|x_t)$ due to the properties of conditioning in mutual informations.

This then defines the active memory utilization rate, leaving the instantaneous predictive capacity to be given by

$$I_X \equiv \lim_{\Delta t\to 0} \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_t)}{p(x_{t+\Delta t})}\right]. \tag{19}$$

Explicitly, the active memory utilization vanishes for Markov, i.e., *memoryless*, processes. Meanwhile, the instantaneous predictive capacity, while not permitting a rate since the numerator encodes path regularity while the denominator does not, will lie in $[0,\infty]$ due to its form as a mutual information.

A. Active memory utilization in continuous time

Here we describe in more detail the behavior and form of the active memory utilization in continuous time which very closely follows the form of the transfer entropy [22]. As with the transfer entropy, while Eq. (18) describes the mean rate of active memory utilization, a local, or pointwise, rate,

$$\dot{m}_X[t, x_{[\tau,t]}] \equiv \lim_{\Delta t\to 0} \frac{1}{\Delta t} \ln \frac{p(x_{t+\Delta t}|x_{[\tau,t]})}{p(x_{t+\Delta t}|x_t)}, \tag{20}$$

is not guaranteed to exist, even if $\dot{M}_X$ does exist, arising if x is anywhere nondifferentiable. Instead, as with the transfer entropy, we must discuss *pathwise* quantities, associated with complete path realizations, in order to associate active memory with individual behavior. As such, we consider the *accumulated* active memory utilization on the interval $[t_0,t]$ to be $M_X^{[t_0,t]}$ which may be written

$$M_X^{[t_0,t]} \equiv \int_{t_0}^{t} dt' \dot{M}_X(t') = \mathbb{E}\big[\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}]\big], \tag{21}$$

where $\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}]$ is the *pathwise active memory utilization* on $[t_0,t]$, defined over complete paths as a logarithmic RN derivative between path measures,

$$\begin{aligned} \mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}] &\equiv \ln \frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]} \\ &\sim \lim_{\Delta t\to 0} \ln \prod_{i=0}^{n} \frac{p(x_{i+1}|x_{\{-k:i\}})}{p(x_{i+1}|x_i)}, \end{aligned} \tag{22}$$

where $t_0 = 0$, $x_i = x_{i\Delta t}$, $n = (t/\Delta t) - 1$, and $k = -\tau/\Delta t$. As with the transfer entropy, the pathwise active memory utilization exists when the relevant RN derivative exists such that $\mathbb{P}_X$ and $\mathbb{P}_X^0$ are absolutely continuous with respect to each other. We denote the dynamics that emerge from the measure $\mathbb{P}_X^0$ the *Markov marginal dynamics*. This in turn leads to the dual definition of the active memory utilization rate

$$\begin{aligned} \dot{M}_X(t) &\equiv \frac{d}{dt}\mathbb{E}\big[\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}]\big] \\ &= \frac{d}{dt}\mathbb{E}\left[\ln \frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]}\right]. \end{aligned} \tag{23}$$

Again, the alternative rate formulation

$$\mathring{M}_X \equiv \lim_{t-t_0\to\infty} \frac{1}{t-t_0}\mathbb{E}\left[\ln \frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]}\right] \tag{24}$$

behaves as $\mathring{M}_X = \dot{M}_X$ when the process is stationary. If the process is both stationary and ergodic then this can be described through the expression

$$\mathring{M}_X = \lim_{t-t_0\to\infty} \frac{1}{t-t_0} \ln \frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]}, \tag{25}$$

which is of use in empirical scenarios where an ensemble of realizations may not be available, but the process can be assumed to be stationary and ergodic.

B. Instantaneous predictive capacity in continuous time

Here we describe the behavior of the instantaneous predictive capacity asymptotically, illustrating this behavior for distinct processes and relating it to the nature of the processes and their sampling paths. If the active information storage characterizes the predictive capacity related to the prediction of some *symbol* x_{t+dt} that is stored in the history of X, the instantaneous predictive capacity characterizes the residual part of this quantity once all predictive capacity related to the prediction of the transition *event* $x_t \to x_{t+dt}$ has been identified (by $\dot{M}_X$). This instantaneous predictive capacity thus accounts for the predictive capacity of the symbol x_{t+dt} "stored" in the current state x_t and as such accounts for the reduction in uncertainty of the state x_{t+dt} given instantaneous properties of the process such as its path regularity. Such a quantity has been defined in such a manner that it does not yield a rate and is thus not a dynamical quantity in the sense of $\dot{M}_X$ or $\dot{T}_{Y\to X}$. A corollary of this is that there exists no such analogous pathwise quantity $\mathcal{I}_X^{[t_0,t]}[x_{[\tau,t]}]$, like in the formulation of the active memory utilization and transfer entropy, stemming from the nonexistence of the proposed measure $\mathbb{P}_X^{\emptyset}$.

Since Markov processes in continuous time possess no active memory utilization, such processes have an instantaneous

predictive capacity equal to their active information storage and thus are ideal for study here. Consequently, an example of instantaneous predictive capacity is the active information storage of the Ornstein-Uhlenbeck process in Eq. (15). We have seen that the active information storage, and thus instantaneous predictive capacity, diverges for such a process, but it is finite when considered as a prediction over some finite time Δt. As such we explore the idea that such a quantity can be characterized by the shape of the function with respect to Δt in the vicinity of $\Delta t \to 0$.

We do so by considering an asymptotic expansion of the instantaneous predictive capacity in the region $\Delta t \to 0$ of the prediction horizon and identify terms in different orders of Δt, similarly to the identification of distinct contributions in [30] (wherein the majority of contributions described behave much like I_X since they cannot be formulated as RN derivatives and thus rates). As such we axiomatically identify these expected components of the instantaneous predictive capacity based on the following relationships,

$$\begin{aligned} I_X^I &\equiv \lim_{\Delta t \to 0} I_X^{(\Delta t)} = I_X, \\ \dot{I}_X^R &\equiv \lim_{\Delta t \to 0} \frac{I_X^{(\Delta t)} - I_X^I}{\Delta t}, \end{aligned} \tag{26}$$

where analogously to $A_X^{(\Delta t)}$ in Eq. (15), we define

$$I_X^{(\Delta t)} \equiv \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_t)}{p(x_{t+\Delta t})}\right]. \tag{27}$$

We denote $\dot{I}_X^R$ the *underlying instantaneous predictive capacity rate* and I_X^I the *nondynamic instantaneous predictive capacity*. Assuming a common general asymptotic form [31] about $\Delta t = 0$,

$$I_X^{(\Delta t)} \sim \exp[-k\Delta t^{-\nu}] \sum_{i=0}^{\infty} \sum_{j=0}^{M(i)} c_{ij} (\ln \Delta t)^j (\Delta t)^{r_i}, \quad t \to 0^+, \tag{28}$$

with $k \geqslant 0$, $\nu > 0$, $r_i \uparrow \infty$, we can then identify contributing components from the asymptotic expansion, which we observe to contribute for $k = 0$, such that we have

$$\begin{aligned} I_X^I &= \lim_{\Delta t \to 0} \sum_{i \forall r_i \leqslant 0} \sum_{j=0}^{M(i)} c_{ij} (\ln \Delta t)^j (\Delta t)^{r_i}, \\ \dot{I}_X^R &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{i \forall r_i > 0} \sum_{j=0}^{M(i)} c_{ij} (\ln \Delta t)^j (\Delta t)^{r_i}. \end{aligned} \tag{29}$$

I_X^I has the same leading order behavior as I_X and characterizes the contribution to the instantaneous predictive capacity not amenable to description as a rate, hence the characterization as an instantaneous, i.e., noninfinitesimal, predictive quantity which has no dynamic analog. $\dot{I}_X^R$ is thus the remaining leading order behavior and characterizes the continuous predictive influence of the process's history in the determination of future states as it dynamically evolves.

These quantities, however, are not guaranteed to be well defined. For instance I_X^I converges iff $c_{ij} = 0 \ \forall \ r_i < 0, j > 0$ and $\dot{I}_X^R$ converges iff $\min_{r_i > 0} r_i = 1$ and $c_{ij} = 0 \ \forall \ r_i > 0, j > 0$. When such conditions are not met, the notion of an instantaneously held contribution and rate become undefined.

It is instructive to examine such contributions for some simple processes. For the Ornstein-Uhlenbeck process detailed above we have $r_i = i$ and the following nonvanishing contributions for $i < 2$,

$$c_{00} = (1/2)\ln[1/(2\kappa)], \quad c_{01} = -\frac{1}{2}, \quad c_{10} = \frac{\kappa}{2}, \tag{30}$$

given by an expansion of Eq. (15), revealing a divergent instantaneous contribution, but a well defined underlying rate since there are no $j > 0$ c_{1j} contributions.

Considering, on the other hand, a process with different path regularity properties, for instance a master equation interpretation of the two species conversion process $A \underset{k_+}{\overset{k_-}{\rightleftharpoons}} B$ [with stationary solution $P_A = k_+/(k_- + k_+)$, $P_B = k_-/(k_- + k_+)$], shown in Appendix A 2, yields

$$\begin{aligned} c_{00} &= -P_A \ln P_A - P_B \ln P_B, \\ c_{10} &= (k_- + k_+)^{-1} k_- k_+ [\ln(k_- k_+) - 2], \\ c_{11} &= 2(k_- + k_+)^{-1} k_- k_+. \end{aligned} \tag{31}$$

Here we find a finite instantaneous contribution (equal to the Shannon information), yet an undefined rate.

We can use these contributions to either assign a limiting instantaneous predictive capacity to each process, in these cases infinity for the Ornstein-Uhlenbeck process and the Shannon entropy for the master equation process, or consider the instantaneous predictive capacity asymptotically and compare the contributions. For instance an Ornstein-Uhlenbeck process with a smaller spring constant κ has an asymptotically faster instantaneous contribution I_X^I, but a smaller underlying rate I_X^R.

It is important to point out that these asymptotic terms demonstrate how I_X is not due to *memory* in any traditional sense, thus lending weight to the nomenclature that we have utilized. If, for example, we take the master equation process associated with Eqs. (31), the rates k_+ and k_- might in reality be the Markov marginal rates of the true rates which may have some deep structure dependent on the past sequence of transitions such that one can predict a *transition* with higher certainty based on this knowledge of the past. I_X cannot detect this dependence and comprises the leading order contribution in A_X. Moreover, it thus follows I_X or A_X would therefore *always* assign larger values to a Markov process without such deep structure merely on the basis of a larger instantaneous Shannon entropy, since the Shannon entropy is $O(1)$ and the part of A_X which can detect long range dependence, $\dot{M}_X dt$, is $O(dt)$.

We point out that the convergence of these instantaneous predictive capacity contributions can be seen to be directly arising from the path regularity of each process. Processes that possess uncertainty in transitions along absolutely continuous sampling paths naturally permit a rate of information "flow" from the history of the process. However, such processes are, by definition, defined in continuous space and possess vanishing uncertainty in the $\Delta t \to 0$ limit due to the absence of discontinuous transitions leading to an unmatched contribution in the form of a differential entropy of a delta

function which diverges. On the other hand, processes with discrete states and sampling paths with a countable number of discontinuities, while similarly attaining vanishing uncertainty along their paths, possess a vanishing conditional Shannon entropy associated with transitions, leading to a well defined instantaneously held contribution. However, the path regularity which affords such a well defined instantaneous contribution directly leads to $\ln(\Delta t)$ contributions arising exactly from the discontinuities, which render the notion of an underlying rate undefined. In both cases the nature of the path regularity leads to logarithmic terms in Δt in either the instantaneous (I_X^I) or rate ($\dot{I}_X^R$) contributions.

V. RELATION TO OTHER INFORMATION-THEORETIC CONCEPTS

A. Discrete time formalisms

It is instructive to construct the quantities analogous to I_X and $\dot{M}_X$ in discrete time and space in order to further illustrate the differences between them and what they quantify. In this case, at time n with time origin 0, we write

$$A_X = I_X + M_X, \quad M_X = \mathbb{E}\left[\ln \frac{p(x_{n+1}|x_{\{0:n\}})}{p(x_{n+1}|x_n)}\right],$$
$$I_X = \mathbb{E}\left[\ln \frac{p(x_{n+1}|x_n)}{p(x_{n+1})}\right], \tag{32}$$

where M_X relates to $\dot{M}_X$ in the same way as the transfer entropy, $T_{Y\to X}$ relates to the transfer entropy rate $\dot{T}_{Y\to X}$. In this case M_X is not a rate, but an $O(1)$ quantity which may be thought of as the active memory utilization associated with the time step $n \to n+1$. If we posit a process $X = x_i$ taking values $x_i \in \mathcal{X} = \{0,1,2,\ldots,N\}$, but at each time step only allow x_i to transition to $x_{i+1} \in \{\{x_i+1, x_i, x_i-1\} \cap \mathcal{X}\}$, this constructs a process with a rudimentary path regularity property, dramatically restricting the space of complete paths $x_{\{0:n\}}$ that are realizable by the process. In this case, because time has been discretized, A_X and I_X are finite because the denominator, $p(x_{n+1})$, *does* produce a probability measure, $\mathbb{P}_X^{\emptyset}$, over paths $x_{\{0:n\}}$; one where each time step is i.i.d. However, while $\mathbb{P}_X$ is absolutely continuous with respect to $\mathbb{P}_X^{\emptyset}$, the two are not equivalent as $\mathbb{P}_X^{\emptyset}$ assigns probability to many paths that $\mathbb{P}_X$ does not, corresponding to paths that the process does not generate. Explicitly, $\mathbb{P}_X^{\emptyset}$ does not account for the property $x_{i+1} \in \{\{x_i+1, x_i, x_i-1\} \cap \mathcal{X}\}$; i.e., it will generate paths that can transition from any part of the phase space to any other in one time step. Consequently, in some steady state such that $p(x_i) > 0 \,\forall\, x_i \in \mathcal{X}$, A_X and I_X get larger as N gets larger without bound, since, as a rough approximation, it is measuring the relative size of $\mathcal{X}$ and $\{\{x_i+1, x_i, x_i-1\} \cap \mathcal{X}\}$. On the other hand, M_X is constructed from measures that agree on which paths are possible and so does not have the unbounded dependence on the size of the state space. As such, one may loosely consider M_X a property of storage associated with paths $x_{\{0:n\}}$, independently of the nature of the ensemble in the wider phase space and its relation to any path regularity, while I_X is a property of storage characterizing precisely this property which we may consider to be the relationship between the ensemble of paths $x_{\{0:n\}}$ and the ensemble of states x_n. In continuous time some path regularity is required for the process to exist and thus the latter component is not expressible as a rate.

B. Information in continuous space and differential entropy

Here we provide an analogy between the issues that we have observed to arise markedly in continuous time and the well known issues surrounding the generalization of Shannon entropy in continuous spaces and attempts to discuss it with differential entropy [32]. Given a continuous space, we may consider the information that can be stored as we increase our ability to resolve the space by partitioning into smaller and smaller regions. As we do this it becomes obvious that the information content of such a variable is, strictly, infinite, following directly from the arbitrary precision with which the variable in question can be specified. This is generally not a useful statement and as such, originally due to Shannon, the notion of differential entropy entered the field without any formal derivation, despite certain (grave) problems associated with it. Indeed it is not clear that such an object has any specific meaning in and of itself. On the other hand, *relative* statements sidestep such issues, are well formed, finite, and are constructed with relative entropies frequently as Kullback-Leibler (KL) divergences. However, it is the exception and not the rule that events in a probability space of some variable are countable and have nonzero probability such that the variable possesses a simple Shannon entropy. As such it is not a particularly demanding claim that if one wants to construct robust, generalizable quantifications of random behavior using information theory (for which we argue computational primitives of stochastic processes should be an example), one should *always* be concerned with how different probability measures relate to each other naturally through KL divergences, which are mathematically underpinned by RN derivatives. For random processes, which are characterized by interrelated collections of random variables or random functions, if one wants to produce dynamic, finite quantities, the appropriate measures must concern the complete paths which the processes generate. An intuitive understanding for this might arise from an appreciation that if one does not compare the full probabilistic behavior of paths over an interval, one will not be meaningfully creating a relative measure that accounts for the additional, and often infinite, precision (and thus information) that is available in the specification of complete random functions, or in the most simple case, in the *timing* of events in continuous time.

It is worth emphasizing this point. Consider, for instance, a special case, where the random process, X, consists of a single impulse (with value $x_{I_t} = 1$, where I_t is the time of the impulse) in a window of length Δt (where $x_t = 0 \,\forall\, t \neq I_t$). The phase space here is discrete, $\{0,1\}$: we cannot distinguish and therefore cannot encode more than $\ln 2$ nats of information into the value of the symbol, but there is further variation that can be exploited due to its timing. The timing can occur at arbitrary precision and so we see that the entire process is functionally identical to a single, continuous, random variable $I_t \in [0, \Delta t]$ with a distribution of behavior entirely captured by a one variable probability density, $p(I_t = t)$. It thus follows that, again, an infinite amount of information could be stored in such a process, indeed for *any* duration Δt. This is an

unavoidable property of the process and, analogously, while one could construct a differential entropy to characterize this distribution, it would, necessarily, inherit all the problems associated with such a quantity in continuous space.

If we examine the form of I_X we can then understand why it is not equipped to balance this differential entropy in the sense of a KL divergence and thus return a convergent relative quantity. The numerator $p(x_{t+dt}|x_t)$ can be iteratively built up into a path probability density functionally identical to $p(I_t = t)$, which can be used to quantify the information content in both the variable x_t *and* its timing. However, information-theoretic quantities based on the denominator, $p(x_t)$, are designed to quantify the information content in the single variable $x_t \in \{0,1\}$, i.e., the nature of the impulse. This form has no ability to relate such an event to the behavior of the system at different times and so is agnostic to the timing of the impulse. This fundamental asymmetry is what causes $\mathcal{I}_X^{[t_0,t]}[x_{[\tau,t]}]$ and a rate of instantaneous predictive capacity to be ill defined and divergent, respectively, a result that can loosely be interpreted as a quantification of the additional (infinite) information content the impulse process can leverage from its timing.

This does *not* mean that quantities that return infinities, such as $\lim_{\Delta t\to 0} A_X/\Delta t$ and $\lim_{\Delta t\to 0} I_X/\Delta t$,[2] are "incorrect" (we emphasize one, in theory, *can* store infinite information in a continuous time process). Rather, in the context of complete paths in continuous time, they are probing answers to relatively unhelpful questions akin to asking the information content of a continuous variable. However, if we change the $p(x_t)$ terms that arise in the construction of these quantities to some other transition probability $p^*(x_{t+\Delta t}|x_t)$, we are creating a relative measure that accounts for, or balances, this infinite precision in the timing, just as KL divergences on continuous spaces account for the infinite precision to which the symbols can be specified. This could correspond to questions such as "how much *more* information can I store using one probability measure, or coding, over another." This has a finite answer and is intimately related to our measure of memory utilization. Again, each individual strategy confers *infinite* information that can be encoded, but there is a finite relative measure.

C. Excess entropy

A well known measure of information storage is the so-called excess entropy [33] which, for stationary processes, is a quantification of the shared or predictive information [34] between the semi-infinite past and future with expression as a mutual information

$$E_X \equiv \lim_{k\to\infty} I[x_{\{n-k;n\}}; x_{\{n+1:n+k+1\}}] \tag{33}$$

in discrete time and

$$E_X \equiv \lim_{r\to\infty} I[x_{[t-r,t]}; x_{(t,t+r]}] \tag{34}$$

in continuous time, where we naturally consider a time origin $\tau \to -\infty$. It should not be underemphasized that the excess entropy possesses analogous properties in continuous time to the active information storage: in general (but not always) it cannot be expressed as an RN derivative between equivalent measures over paths. This can be seen by understanding that the excess entropy contains the active information storage. For instance, it is known that, for a stationary process, in discrete time, the active information storage relates to the excess entropy as per [4,6]

$$E_X = \sum_{k=0}^{\infty} \left(A_X - A_X^{(k)}\right) = A_X + \sum_{k=1}^{\infty} \left(A_X - A_X^{(k)}\right), \tag{35}$$

where

$$A_X^{(k)} \equiv \mathbb{E}\left[\ln \frac{p(x_{n+1}|x_{\{n-k+1:n\}})}{p(x_{n+1})}\right], \tag{36}$$

with $A_X^{(k)} \leqslant A_X^{(k+1)}$ and $A_X^{(0)} = 0$. In continuous time an analogous relation holds,

$$E_X \equiv A_X + \int_0^{\infty} ds\ \Delta\dot{A}_X^{(s)}, \tag{37}$$

where

$$\Delta\dot{A}_X^{(s)} \equiv \lim_{\Delta t\to 0} \frac{1}{\Delta t}\left(A_X - A_X^{(s)}\right) \geqslant 0,$$
$$A_X^{(s)} \equiv \mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[t-s,t]})}{p(x_{t+\Delta t})}\right]. \tag{38}$$

Consequently, where A_X is divergent, it follows that E_X is too. Note, $\Delta\dot{A}_X^{(s)}$ is $O(1)$, despite neither A_X or $A_X^{(s)}$ individually leading to a well defined rate. This can be seen by noting that this expression is identical to

$$E_X = A_X + \int_0^{\infty} ds\ \Delta\dot{M}_X^{(s)}, \tag{39}$$

where $\Delta\dot{M}_X^{(s)} = \Delta\dot{A}_X^{(s)}$, i.e.,

$$\Delta\dot{M}_X^{(s)} \equiv \dot{M}_X - \dot{M}_X^{(s)},$$
$$\dot{M}_X^{(s)} \equiv \lim_{\Delta t\to 0} \frac{1}{dt}\mathbb{E}\left[\ln \frac{p(x_{t+\Delta t}|x_{[t-s,t]})}{p(x_{t+\Delta t}|x_t)}\right], \tag{40}$$

with both $\dot{M}_X$ and $\dot{M}_X^{(s)}$ expected to be convergent rates since they lead to integrated RN derivatives and where, typically, $\lim_{s\to\infty} \dot{M}_X^{(s)} = \dot{M}_X$ such that $\int_0^{\infty} ds\ \Delta\dot{M}_X^{(s)} \in [0,\infty]$. Consequently, while it may be common to observe $E_X = \infty$ and $A_X = \infty$ (e.g., for continuous processes such as the Ornstein-Uhlenbeck process and generalizations), their difference, $E_X - A_X$, may yet be a convergent $O(1)$ quantity. And importantly, if it does not converge, it relates to the consideration of an infinite interval and behavior in $\dot{M}_X^{(s)}$ which causes the appropriate integral on $[0,\infty]$ to be divergent, not the nonequivalence or nonexistence of the measures under consideration.

Somewhat reminiscent of the discussion in Sec. V B, again, we see a situation in which two information-theoretic measures may be infinite, but possess a *relative* difference that is convergent. Explicitly, both E_X and A_X can be infinite, but possess a finite difference in the limit. Indeed, since $A_X = I_X + \dot{M}\Delta t$

[2]We note the quantities A_X and I_X may be finite, and indeed meaningful, even if $A_X/\Delta t$ or $I_X/\Delta t$ are not.

this may, again in the limit, be written

$$\lim_{\Delta t\to 0} E_X - A_X = \lim_{\Delta t\to 0} E_X - I_X = \int_0^\infty ds\ \Delta\dot{M}_X^{(s)}, \quad (41)$$

implying that, more specifically, whenever I_X diverges, E_X does also such that a finite E_X implies a finite (or vanishing) I_X. Further, the quantity in Eq. (41) exists in the literature as the so-called "elusive" information [35], $\sigma_X \equiv \lim_{s\to\infty} I[x_{(t,t+s]}; x_{[t-s,t)}|x_t]$, since

$$\begin{aligned}
&\int_0^\infty ds\ \Delta\dot{M}_X^{(s)} \\
&= \lim_{r\to\infty}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_0^r ds\ \mathbb{E}\left[\ln\frac{p(x_{t+\Delta t}|x_{[t-r,t]})}{p(x_{t+\Delta t}|x_{[t-s,t]})}\right] \\
&\simeq \lim_{r\to\infty}\lim_{\Delta t\to 0}\sum_{i=0}^{n=r/\Delta t}\mathbb{E}\left[\ln\frac{p(x_{t+\Delta t}|x_{[t-r,t]})}{p(x_{t+\Delta t}|x_{[t-i\Delta t,t]})}\right] \\
&\simeq \lim_{r\to\infty}\lim_{\Delta t\to 0}\sum_{i=0}^{n=r/\Delta t}\mathbb{E}\left[\ln\frac{p(x_{t+(i+1)\Delta t}|x_{[t-r,t+i\Delta t]})}{p(x_{t+(i+1)\Delta t}|x_{[t,t+i\Delta t]})}\right] \\
&= \lim_{r\to\infty}\mathbb{E}\left[\ln\frac{d\mathbb{P}_X[x_{(t,t+r]}|x_{[t-r,t]}]}{d\mathbb{P}_X[x_{(t,t+r]}|x_t]}\right]. \quad (42)
\end{aligned}$$

Here we have first represented the integral as a sum on line 2 which converges simultaneously with the vanishing prediction horizon. Then, by assuming stationarity and time homogeneity, we have moved to line three by relabeling the time indexes in the probabilities forward in time by $i\Delta t$. By taking the sum inside the logarithm as a product, the sequence then converges to an RN derivative between the constructed paths in the limit in Δt, taken on the last line. This quantity, being a logarithm of an RN derivative between equivalent measures, is $O(1)$. This, in turn, clarifies and emphasizes a different approach to information-theoretic quantities in continuous time as opposed to that offered in, for example, Ref. [30]. In our approach one does not simply scale *all* quantities one might consider by a time discretization parameter, but instead identifies rates and integrated quantities, treating such quantities differently. The elusive information, like the expectation of the pathwise active memory utilization and pathwise transfer entropy, is an integrated quantity associated with complete paths and as such is not meaningfully expressed as a rate with respect to a small time discretization since it concerns behavior that persists far beyond any such timescale. On the other hand, if the integral that characterizes the elusive information does not converge, such that $\lim_{dt\to 0} E_X - I_X = \infty$, one can construct a rate, in the alternative sense, with respect to the entire process by defining, and considering in the limit $t\to\infty$,

$$\mathring{\sigma}_X(t) = \lim_{r\to\infty}\frac{1}{t} I\left[x_{(t_0,t_0+t]}; x_{[t_0-r,t_0)}\middle|x_{t_0}\right], \quad (43)$$

or perhaps, when considering the explicit growth of the expected pathwise quantity that underlies σ_X,

$$\begin{aligned}
\dot{\sigma}_X(t) &= \lim_{r\to\infty}\frac{d}{dt} I\left[x_{(t_0,t_0+t]}; x_{[t_0-r,t_0)}\middle|x_{t_0}\right] \\
&= \dot{M}_X - \dot{M}_X^{(t)}, \quad (44)
\end{aligned}$$

noting that these quantities are not equivalent. Nowhere, however, is the quantity $\lim_{\Delta t\to 0}\sigma_X/\Delta t$ implicated in the construction of such rates or expected to converge—for precisely the same reason it is not expected to for the quantities $T_{Y\to X}^{[t_0,t]}/\Delta t$ or $M_X^{[t_0,t]}/\Delta t$. Analogously, it follows that E_X, A_X, and I_X are natural $O(1)$ quantities (though they may be infinite) and as such neither should we consider $\lim_{\Delta t\to 0} E_X/\Delta t$, etc.

VI. COMPONENTS OF INFORMATION STORAGE IN JUMP AND NEURAL SPIKING PROCESSES

With the preceding measures of instantaneous predictive capacity and active memory utilization set out, we can describe such quantities in specific systems such that, when complemented by previous work on transfer entropy, a complete picture of information processing, as understood by the complementary description of memory and signaling, can be described. We acknowledge the parallel description of such processes, in the absence of extrinsic processes, in [36–39], which ultimately can be viewed as complementary in the sense of the connection between excess entropy and the quantities we consider as per Sec. V C.

A. Jump processes

In previous work we described how to construct the relevant pathwise transfer entropy functional for jump processes for which neural spiking processes are a specific example [22]. Much of the resulting structure for active memory utilization is analogous, but we reiterate the key points. We imagine, for simplicity, a discrete state space $x\in\mathcal{X}$. These are then stochastic processes characterized by intermittent transitions between the states in $\mathcal{X}$ and where the states are constant in between these transitions. As such a path on the interval $[t_0,t]$, $x_{[t_0,t]}$, is characterized by the start and end times t_0 and t, its starting state x_0, and the times t_i it transitions into the states x_i such that we write $x_{[t_0,t]}\equiv\{t,\{t_i,x_i\}_0^{N_x}\}$ where $\{t_i,x_i\}_0^{N_x}\equiv\{t_0,x_0,t_1,x_1,\ldots,t_{N_x},x_{N_x}\}$ and where N_x is the number of transitions in x on the interval.

Any given path realization then possesses a probability density [40], constructed with the entire knowledge of the history of x, given some time origin τ, at each point in time

$$p_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right] = \left(\prod_{i=1}^{N_x} W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]\right)\exp\left[-\int_{t_0}^t \lambda_X[x_{[\tau,t)}]\right] \quad (45)$$

and a probability density constructed with knowledge only of the current value of x at each point in time

$$p_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right] = \left(\prod_{i=1}^{N_x} W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]\right)\exp\left[-\int_{t_0}^t \lambda_X^0[x_t^-]\right], \quad (46)$$

where W are transition rates, λ are escape rates, and $x_t^- = \lim_{t'\nearrow t} x_{t'}$ such that

$$\begin{aligned}
W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right] &= \lim_{\Delta t\to 0}\frac{1}{dt}p\left(x_{t_i}^-\to x_{t_i}\in[t_i,t_i+\Delta t]|x_{[\tau,t_i)}\right), \\
W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right] &= \lim_{\Delta t\to 0}\frac{1}{dt}p\left(x_{t_i}^-\to x_{t_i}\in[t_i,t_i+\Delta t]|x_{t_i}^-\right),
\end{aligned}$$

$$\lambda_X[x_{[\tau,t)}] = \sum_{x' \neq x_t^- \in \mathcal{X}} W_X[x'|x_{[\tau,t)}],$$
$$\lambda_X^0[x_t^-] = \sum_{x' \neq x_t^- \in \mathcal{X}} W_X^0[x'|x_t^-]. \quad (47)$$

We note that such quantities may depend on the time t_i or t respectively should the process be nonstationary. The RN derivative that constitutes the pathwise active memory utilization is then the ratio of these two quantities such that

$$\begin{aligned}\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}] &= \ln \frac{d\mathbb{P}_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]} \\ &= \ln \frac{p_X\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]dt_1 \ldots dt_{N_x}}{p_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]dt_1 \ldots dt_{N_x}} \\ &= \sum_{i=1}^{N_x} \ln \frac{W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]}{W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]} \\ &\quad - \int_{t_0}^{t} \left(\lambda_X[x_{[\tau,t')}] - \lambda_X^0[x_{t'}^-]\right)dt'. \end{aligned} \quad (48)$$

As with the transfer entropy, there is a continuous integral component related to the waiting times between transitions and N_x instantaneous contributions due to transitions between states. These instantaneous jumps are, in this instance, what stop a local rate [in the form of Eq. (20)] from being well defined. Consequently we may decompose the total change of $\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}]$ with time into these two components related to transitions $\Delta\mathcal{M}_X^t$ and waiting times $\dot{\mathcal{M}}_X^{nt}$, the latter of which *does* permit a rate, such that

$$\begin{aligned}\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}] &= \sum_{i=1}^{N_x} \Delta\mathcal{M}_X^t(t_i) + \int_{t_0}^{t} dt'\, \dot{\mathcal{M}}_X^{nt}(t'), \\ \Delta\mathcal{M}_X^t(t_i) &= \ln \frac{W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]}{W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]}, \\ \dot{\mathcal{M}}_X^{nt}(t) &= \lambda_X^0[x_{t'}^-] - \lambda_X[x_{[\tau,t')}]. \end{aligned} \quad (49)$$

Importantly, since $\lambda_X^0[x_t^-]$ is just a marginalized average of $\lambda_X[x_{[\tau,t')}]$, when the expectation is taken we find

$$\begin{aligned}\mathbb{E}\left[\dot{\mathcal{M}}_X^{nt}(t)\right] &= \mathbb{E}\left[\lambda_X[x_{[\tau,t)}] - \lambda_X^0[x_t^-]\right] \\ &= 0. \end{aligned} \quad (50)$$

Consequently we may write

$$\mathbb{E}\left[\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}]\right] = \mathbb{E}\left[\sum_{i=1}^{N_x} \ln \frac{W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]}{W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]}\right], \quad (51)$$

and thus

$$\begin{aligned}\dot{M}_X(t) &= \frac{d}{dt}\mathbb{E}\left[\sum_{i=1}^{N_x} \ln \frac{W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]}{W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]}\right] \\ &= \mathbb{E}\left[\left(1 - \delta_{x_t,x_t^-}\right) \ln \frac{W_X[x_t|x_{[\tau,t)}]}{W_X^0[x_t|x_t^-]}\right], \end{aligned} \quad (52)$$

where δ_{x_t,x_t^-} is the Kronecker delta. The alternative rate is given by

$$\mathring{M}_X = \lim_{t-t_0\to\infty} \frac{1}{t-t_0}\mathbb{E}\left[\sum_{i=1}^{N_x} \ln \frac{W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]}{W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]}\right] \quad (53)$$

equal to $\dot{M}_X$ when the process is stationary and we may write

$$\mathring{M}_X = \lim_{t-t_0\to\infty} \frac{1}{t-t_0}\sum_{i=1}^{N_x} \ln \frac{W_X\left[x_{t_i}\middle|x_{[\tau,t_i)}\right]}{W_X^0\left[x_{t_i}\middle|x_{t_i}^-\right]} \quad (54)$$

when the process is ergodic.

On the other hand, when considering the instantaneous predictive capacity, we emphasize, no analogous pathwise quantity $\mathcal{I}_X^{[t_0,t]}[x_{[\tau,t]}]$, and as discussed, no rate, exists for a direct comparison. However the mean rates $\dot{T}_{Y\to X}$ and $\dot{M}_X$ that emerge from this description sit alongside the asymptotic contributions to I_X. These contributions are obtained in Appendix A 2 and, for X taking values in a set of discrete states $\mathcal{X}$, are given by

$$\begin{aligned} c_{00} &= -\sum_{x_t\in\mathcal{X}} P(x_t)\ln P(x_t), \\ c_{10} &= \sum_{x_t^-\in\mathcal{X}} \sum_{x_t\neq x_t^-\in\mathcal{X}} P(x_t^-)W_X^0[x_t|x_t^-] \\ &\quad \times \left[\ln \frac{W_X^0[x_t|x_t^-]P(x_t^-)}{P(x_t)} - 1\right], \\ c_{11} &= \sum_{x_t^-\in\mathcal{X}} \sum_{x_t\neq x_t^-\in\mathcal{X}} P(x_t^-)W_X^0[x_t|x_t^-], \end{aligned} \quad (55)$$

which is merely a generalization of Eq. (31). As such we acknowledge the limiting value $I_X(t)$ is given by the Shannon entropy of the system at the time t.

B. Neural spiking processes

We consider an idealization of neural processes whereby a realization of the process consists entirely of indistinguishable, nonoverlapping, events (spikes) of duration 0 seconds, such that any given path is characterized entirely by the timings of such events, $t_{i+1} > t_i$, etc., i.e., a point process. This can be constructed, in the sense of a stochastic process detailed above, in a number of ways, but here, rather than follow [22] where X concerned the number of spikes that occurred since a time origin, we instead consider the limit of a two state system with the states, 0 and 1, corresponding to "not spiked" and "spiked" respectively in the manner of burst noise or a telegraph process. Since there are only two states we have $W_X[x_{t_i}|x_{[\tau,t_i)}] = \lambda_X[x_{[\tau,t_i)}]$ and $W_X^0[x_{t_i}|x_{t_i}^-] = \lambda_X^0[x_{t_i}^-]$. Finally, to achieve the reduction to a point process, we let $\lambda_X[x_{[\tau,t_i)}]_{x_{t_i}^-=1} = \lambda_X^0[x_{t_i}^- = 1]$ such that there is no non-Markov dependence in the transition that characterizes return to the unspiked state and then we consider the limit $\lambda_X^0[x_{t_i}^- = 1] \to \infty$ such that the transition from the spiked state to the unspiked state is immediate. These two conditions ensure that there is no contribution to the active memory utilization due to return

to the unspiked state following spikes since

$$\ln \frac{\lambda_X\left[x_{[\tau,t_i)}\right]_{x_{t_i}^-=1}}{\lambda_X^0\left[x_{t_i}^-=1\right]} = 0 \tag{56}$$

and that there is no contribution to the integral component from being in the spiked state since as $\lambda_X^0[x_{t_i}^- = 1] \to \infty$, x is in state 0 with probability 1, such that

$$\int_{t_0}^{t} (\lambda_X[x_{[\tau,t')}] - \lambda_X^0[x_{t'}^-])dt' = \int_{t_0}^{t} (\lambda_X[x_{[\tau,t')}]_{x_{t'}^-=0} - \lambda_X^0[x_{t'}^- = 0])dt'. \tag{57}$$

In this limit we may then characterize the process with a single transition or escape rate $\lambda_X[x_{[\tau,t)}]$, with expectation value $\lambda_X^0[x_t^-] = \lambda_X^0(t)$ which is understood to be the conditional spike rate with knowledge of the history of x and the mean spike rate, respectively. In addition, since the return transitions happen instantaneously we may now simply characterize the paths with the times of the spike events $x_{[t_0,t]} \equiv \{t, \{t_i\}_0^{N_x}\}$.

Consequently, for neural spike processes we find

$$\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}] = \sum_{i=1}^{N_x} \ln \frac{\lambda_X\left[x_{[\tau,t_i)}\right]}{\lambda_X^0(t_i)} - \int_{t_0}^{t} (\lambda_X[x_{(\tau,t')}] - \lambda_x^0(t'))dt', \tag{58}$$

which in turn possesses mean rates

$$\mathring{M}_X = \lim_{t-t_0\to\infty} \frac{1}{t-t_0} \mathbb{E}\left[\sum_{i=1}^{N_x} \ln \frac{\lambda_X\left[x_{[\tau,t_i)}\right]}{\lambda_X^0(t)}\right],$$
$$\dot{M}_X(t) = \mathbb{E}\left[\left(1-\delta_{x_t,x_t^-}\right) \ln \frac{\lambda_X[x_{[\tau,t)}]}{\lambda_X^0(t)}\right], \tag{59}$$

both equal for stationary processes and equal to

$$\mathring{M}_X = \lim_{t-t_0\to\infty} \frac{1}{t-t_0} \sum_{i=1}^{N_x} \ln \frac{\lambda_X\left[x_{[\tau,t_i)}\right]}{\lambda_X^0} \tag{60}$$

when the process is also ergodic.

The instantaneous predictive capacity for spike processes, like with the active memory utilization, is simply a special case of that for jump processes. Indeed, as formulated here, it is a special case of the two species conversion process in Sec. IV B with A corresponding to the unspiked state and B corresponding to the spiked state with rates $k^+ = \lambda_X^0$ and $k^- = \mu$ in the limit of μ being taken to infinity such that we have

$$c_{00} = \lim_{p\to 1} -p\ln p - (1-p)\ln(1-p) = 0,$$
$$c_{10} = \lim_{\mu\to\infty} \left[\lambda_X^0(t)+\mu\right]^{-1} \lambda_X^0(t)\mu\left\{\ln\left[\lambda_X^0(t)\mu\right] - 2\right\} = \infty,$$
$$c_{11} = \lim_{\mu\to\infty} 2\left[\lambda_X^0(t)+\mu\right]^{-1} \lambda_X^0(t)\mu = 2\lambda_X^0(t). \tag{61}$$

As such we acknowledge the limiting value $I_X(t) = 0$, noting that this can only emerge here as a consequence of assigning 0 measure to the spiking phenomena on such timescales despite the fact that they occur almost surely in a sufficiently long time interval, and a divergent underlying rate, dominated by terms linear in the mean Markov intensity of the process. Indeed this limiting value, $I_X = 0$, in conjunction with the corollary of Eq. (41) that finite excess entropy, E_X, implies finite (or vanishing) instantaneous predictive capacity, contextualizes other results reporting convergent excess entropy in the case of point processes [37,39].

A note of caution with respect to empirical estimation techniques

As with the estimation of the transfer entropy in such a setting [22], we anticipate that the most efficient estimation of $\dot{M}_X$ in this context will emerge when utilizing an estimator designed specifically to utilize interspike time intervals as relevant continuous variables. However, as with the transfer entropy a time binned approximation, while perhaps inefficient, will, in theory, be able to capture, in the limit, the behavior of the above formalism. It is here, however, that spike train data can appear to be uniquely ambiguous when discretized in this fashion and for which we anticipate possible confusion if not performed carefully. For instance, if one attempts to discretize such that given a time interval $[0,t]$ with N_x spikes based on a time resolution of Δt, such that one creates N_x "spiked" bins and $(t/\Delta t) - N_x$ "not spiked" bins, one might appear to observe a convergent A_X "rate" and vanishing I_X "rate" by utilizing $p(\text{spike}) = N_x\Delta t/t$, calculating the relevant quantities, and *then* taking the limit. But this is incorrect as it is conflating $p(x_{t+dt})$ and $p(x_{t+dt}|x_t)$. Intuition as to the origin of the error in this hypothetical scheme can be gained by realizing that there is no discretization resolution, Δt, where a spike event takes up any more than one bin. Consequently *all* finite probability associated with the spiked state is entirely an artifact of the discretization procedure reflecting the correct probability $p(\text{spike}) = 0$. That is, such a discretization procedure would be conflating the "state" of being spiked with associated vanishing probability, with the more appropriate characterization of a spike as a *transition* with associated probability density, with respect to a vanishing time interval; i.e., it would be interpreting $p(x_t)$ as a probability density, when it is not. As such, the quantity $N_x\Delta t/t$ reflects the probability of a spike any time *within* an interval Δt, based on the history free statistics of the entire interval $[0,t]$, i.e., $\lambda_X^0\Delta t$ as per the definition of the transition (spike) rates in Eqs. (47), which reflects the behavior of $p(x_{t+dt}|x_t)$. Replacing $p(x_{t+dt})$ with $p(x_{t+dt}|x_t)$ in A_X returns $\dot{M}_X dt$ and thus the hypothetical naive calculations of a rate of A_X would rather have been approximations of $\dot{M}_X$.

C. Simple analytical examples

1. Stationary Poisson process with refractory period

Perhaps the most simple, nontrivial, process X for which active memory utilization is present in such a setting is a simple Poisson model of spiking with the introduction of a refractory period following any spike, of duration Δ_x seconds, during which the process cannot subsequently spike again. Defining t_x to be the time of the most recent spike in X, and thus $t^x = t - t_x$ as the time *since* the last spike in X, we may specify this through the conditional spike rate

$$\lambda_X[x_{[\tau,t)}] = \lambda_X(t^x) = \begin{cases} \mu, & t^x \geqslant \Delta_x, \\ 0, & t^x < \Delta_x. \end{cases} \tag{62}$$

Calculation of the relevant quantities can be achieved by exploiting the fact that the process is piecewise Markovian between refractory periods and must appear Markovian and stationary when constructing the measure $\mathbb{P}_X^0$. Consequently, when we calculate λ_X^0 we can simply recognize that the characteristic time frame required to achieve a single spike following any previous spike is simply $\Delta_x + \mu^{-1}$ such that we have

$$\lambda_X^0 = \frac{\mu}{1+\mu\Delta_x}. \tag{63}$$

Next we can then use these expressions to calculate the active memory utilization rate contribution *per spike*, due to that spike varying only in its timing since the previous spike, given by

$$\begin{aligned}
&\lim_{t\to\infty} \int_0^t \exp\left[-\int_{t_x}^{t'} \lambda_X(t'')dt''\right] \lambda_X(t') \ln \frac{\lambda_X(t')}{\lambda_X^0} dt' \\
&= \int_{\Delta_x}^{\infty} \exp\left[-\mu(t-\Delta_x)\right] \mu \ln(1+\mu\Delta_x) dt \\
&= \ln(1+\mu\Delta_x).
\end{aligned} \tag{64}$$

There are then λ_X^0 of these spike events per unit time, each with the average contribution above. Further, since the contribution for nonspiking behavior must vanish on average due to Eq. (50), it then follows that the active memory utilization rate is

$$\dot{M}_X = \frac{\mu \ln(1+\mu\Delta_x)}{1+\mu\Delta_x}. \tag{65}$$

We see that μ serves, primarily, to control the number of spikes per unit time, thus scaling the number of prediction events and therefore the total scale of the active memory utilization rate as a simple prefactor in addition to terms in $1+\mu\Delta_x$. Interestingly, however, this rate exhibits a maximum with respect to Δ_x for any given μ corresponding to $\Delta_x = \Delta_x^{\max} = (e-1)/\mu$ such that the largest active memory utilization rate is given by $\dot{M}_X^{\max} = \mu/e$. Note that the form of these expressions, dependent on e, is not due to the choice of base of the logarithm used here. The existence of this maximum arises through the balance of two factors: increasing Δ_x increases the contribution *per spike* through the logarithmic term as per Eq. (64), but also reduces the total number of expected spikes per unit time and thus total rate through the inverse $1+\mu\Delta_x$ term. We may also understand this phenomena through the distinguishability of the processes that correspond to $\mathbb{P}_X$ and $\mathbb{P}_X^0$. For values $\Delta_x \ll \Delta_x^{\max}$, the refractory period is not meaningfully impacting the dynamics such that both appear Poissonian. On the other hand, values $\Delta_x \gg \Delta_x^{\max}$ also make the two processes, on aggregate, appear similar since, for the increasingly representative majority of the process (the refractory period), they behave as two processes with an arbitrarily low spike rate, despite the increase in mean contribution per spike.

Finally, as a spiking process, with a two value state space with vanishing Shannon entropy, the behavior of I_X is given by Eq. (61).

2. *Nonstationary event driven spiking process*

To illustrate the form of the active memory utilization in a marginally more complicated setting we use an elaborated simple toy model consisting of two spiking processes, X and Y. We specify Y to be a *deterministic* spike train which spikes regularly with period Δ_y. Noting that we have set the time origin $\tau = t_0$ for convenience, this means that the process always realizes the specific path $y_{[t_0,t)} = y^*_{[t_0,t)} = \{\ldots, -\Delta_y, 0, \Delta_y, 2\Delta_y, \ldots\}$, designed in this way to induce nonstationary behavior in X. In this way, Y could be considered to be some external stimuli occurring at regular time intervals, triggering separate trials in the sense of an event driven neural experiment. Next we specify that X has a probability, c, of spiking within Δ_x seconds of the spike in Y and that the spike occurs with uniform distribution on the interval $[t_y, t_y+\Delta_x]$ where $t_y \leqslant t$ is the time of the most recent spike in Y. Outside of this window X cannot spike. We also insist on a refractory period of Δ_x seconds in the X neuron during which it cannot spike immediately after spiking. Finally, we also specify that $2\Delta_x < \Delta_y$ such that there can be no ambiguity in which spike in Y is responsible for the possible spike in X and only one spike in X can result from any given spike in Y. A conditional spike rate that achieves this is given by

$$\begin{aligned}
&\lambda_{X|Y}\left[x_{[t_0,t)}, y_{[t_0,t)}\right] \\
&= \lambda_{X|Y}(t_x, t_y, t) \\
&= \begin{cases} \frac{c}{\Delta_x - c(t-t_y)}, & t < t_y + \Delta_x \text{ and } t \geqslant t_x + \Delta_x, \\ 0, & t \geqslant t_y + \Delta_x \text{ or } t < t_x + \Delta_x, \end{cases}
\end{aligned} \tag{66}$$

where $t_x \leqslant t$ is the time of the last spike in x. This is easily observed since the probability of an absence of spikes on $[t_y, t]$, $t < t_y + \Delta_x$, is given by

$$\exp\left[-\int_{t_y}^{t} dt' \frac{c}{\Delta_x - c(t'-t_y)}\right] = 1 - \frac{c(t-t_y)}{\Delta_x}, \tag{67}$$

such that the probability density of spiking at time t is given by

$$p(t) = \frac{d}{dt}\frac{c(t-t_y)}{\Delta_x} = \frac{c}{\Delta_x}, \tag{68}$$

i.e., a uniform distribution. Note that $\lambda_{X|Y}$ can be rewritten entirely in terms of the variables $t^x = t - t_x$ and $t^y = t - t_y$, the times *since* the most recent spike in X and Y, i.e., from variables encoded into the *sequences* of the path histories of X and Y, respectively, independently of the time, indicating that the behavior encoded by $\lambda_{X|Y}$ is time homogeneous. Since Y is deterministic, the form of the conditional spike rate in terms of X is trivial. Noting the shorthand

$$\begin{aligned}
&\int dy_{[t_0,t)} p\left[x_{[t_0,t)}, y_{[t_0,t)}\right] \\
&\equiv p_{N_x,0}\left(t, \{t^x\}_0^{N_x}\right) + \int_{t_0}^{t} dt_1^y \; p_{N_x,1}\left(t, \{t^x\}_0^{N_x}, t_0^y\right) \\
&\quad + \int_{t_0}^{t} dt_1^y \int_{t_1^y}^{t} dt_2^y \; p_{N_x,2}\left(t, \{t^x\}_0^{N_x}, \{t^y\}_0^2\right) \\
&\quad + \sum_{i=3}^{\infty} \int_{t_0}^{t} dt_1^y \ldots \int_{t_{i-1}^y}^{t} dt_i^y \; p_{N_x,i}\left(t, \{t_x\}_0^{N_x}, \{t_y\}_0^i\right),
\end{aligned} \tag{69}$$

where $x_{[t_0,t)} \equiv \{t, \{t^x\}_0^{N_x}\}$, a path consisting of N_x spikes, $p_{N_x,i}$ is the probability density of a joint spike sequence on $[t_0, t)$

with such N_x spikes in X and i spikes in Y, where t_x^i and t_y^i are the times of the ith spikes in X and Y, respectively, and $t_0^x = t_0^y = t_0$, we may write

$$\begin{aligned}
&\lambda_X\left[x_{[t_0,t)}\right] \\
&\quad= \frac{1}{p\left[x_{[t_0,t)}\right]} \int dy_{[t_0,t)} \lambda_{X|Y}\left[x_{[t_0,t)}, y_{[t_0,t)}\right] p\left[x_{[t_0,t)}, y_{[t_0,t)}\right] \\
&\quad= \int dy_{[t_0,t)} \lambda_{X|Y}\left[x_{[t_0,t)}, y_{[t_0,t)}\right] \frac{p\left[x_{[t_0,t)}\middle| y_{[t_0,t)}\right] p\left[y_{[t_0,t)}\right]}{p\left[x_{[t_0,t)}\right]} \\
&\quad= \int dy_{[t_0,t)} \lambda_{X|Y}\left[x_{[t_0,t)}, y_{[t_0,t)}\right] \\
&\qquad\times \frac{p\left[x_{[t_0,t)}|y_{[t_0,t)}\right]}{p\left[x_{[t_0,t)}\right]} \delta\left(y_{[t_0,t)} - y^*_{[t_0,t)}\right) \\
&\quad= \lambda_{X|Y}\left[x_{[t_0,t)}, y^*_{[t_0,t)}\right] \frac{p\left[x_{[t_0,t)}|y^*_{[t_0,t)}\right]}{p\left[x_{[t_0,t)}\right]} \\
&\quad= \lambda_{X|Y}\left[x_{[t_0,t)}, y^*_{[t_0,t)}\right],
\end{aligned} \tag{70}$$

since $p[x_{[t_0,t)}] = \int dy_{[t_0,t)} p[x_{[t_0,t)}|y_{[t_0,t)}]\delta(y_{[t_0,t)} - y^*_{[t_0,t)}) = p[x_{[t_0,t)}|y^*_{[t_0,t)}]$. That is, the joint process with a stationary dependence in X on a deterministic Y appears as a *nonstationary* process in X alone, such that

$$\begin{aligned}
\lambda_X\left[x_{[t_0,t)}\right] &= \lambda_X(t_x, t) \\
&= \begin{cases} \frac{c}{\Delta_x - c(t - n\Delta_y)}, & t \in [n\Delta_y, n\Delta_y + \Delta_x] \\ & \text{and } t \geqslant t_x + \Delta_x, \\ 0, & t \notin [n\Delta_y, n\Delta_y + \Delta_x] \\ & \text{or } t < t_x + \Delta_x, \end{cases}
\end{aligned} \tag{71}$$

with $n \in \mathbb{Z}$. Note, in contrast to $\lambda_{X|Y}$, which can be written in terms of t^x and t^y, λ_X cannot be written solely in terms of t^x and thus retains dependence on the time t, reflecting induced time inhomogeneity, leading to the observed nonstationary behavior. This nonstationary behavior is carried into $\lambda_X^0(t)$ which is given by considering the probability of a single spike in $[n\Delta_y, (n+1)\Delta_y]$, such that we need only consider $t_x \leqslant t - \Delta_x$, viz.,

$$\begin{aligned}
\lambda_X^0(t) &= \frac{\Delta_x - c(t - n\Delta_y)}{\Delta_x} \lambda_X(t_x, t \in [n\Delta_y, n\Delta_y + \Delta_x]) \\
&\quad + c\lambda_X(t_x, t \notin [n\Delta_y, n\Delta_y + \Delta_x]) \\
&= \begin{cases} \frac{c}{\Delta_x}, & t \in [n\Delta_y, n\Delta_y + \Delta_x], \\ 0, & t \notin [n\Delta_y, n\Delta_y + \Delta_x], \end{cases}
\end{aligned} \tag{72}$$

such that the Markov marginal process cannot determine whether the process is in its refractory period or not. Since the process is nonstationary, it follows that $\dot{M}_X$ is not a constant here. Recognizing, by construction, that we have a periodic process Y, an at most one to one mapping between spikes in Y and X, and the property that outside of the refractory period, X is piecewise Markov, we may, without loss of generality, consider times $n\Delta_y \leqslant t < (n+1)\Delta_y$, treating $n\Delta_y$ as an effective time origin. Then, due to the refractory period, there is only a nonzero contribution for spike histories that have 0 previous spikes in $[n\Delta_y, t]$, thus, again by construction, satisfying $t_x \leqslant t - \Delta_x$. Consequently, we may write

$$\begin{aligned}
\dot{M}_X(t) &= \mathbb{E}\left[\left(1 - \delta_{x_t, x_t^-}\right) \ln \frac{\lambda_X\left[x_{[t_0,t)}\right]}{\lambda_X^0(t)}\right] \\
&= \exp\left[-\int_{n\Delta_y}^{t} dt'\, \lambda_X(t_x \leqslant t - \Delta_x, t')\right] \\
&\quad\times \lambda_X(t_x \leqslant t - \Delta_x, t) \ln \frac{\lambda_X(t_x \leqslant t - \Delta_x, t)}{\lambda_X^0(t)} \\
&= \begin{cases} \frac{c}{\Delta_x} \ln \frac{\Delta_x}{\Delta_x - c(t - n\Delta_y)}, & t \in [n\Delta_y, n\Delta_y + \Delta_x], \\ 0, & t \notin [n\Delta_y, n\Delta_y + \Delta_x], \end{cases}
\end{aligned} \tag{73}$$

recovering the general case by once again letting $n \in \mathbb{Z}$. Further, we may consider $\mathring{M}_X$, which in this instance is simply

$$\begin{aligned}
\mathring{M}_X &= \lim_{t\to\infty} \frac{1}{t} \int_0^t \dot{M}_X(t')dt' = \frac{1}{\Delta_y} \int_{n\Delta_y}^{(n+1)\Delta_y} \dot{M}_X(t)dt \\
&= \Delta_y^{-1}[c + (1-c)\ln(1-c)].
\end{aligned} \tag{74}$$

Δ_y serves simply to control the flow of predictable events and thus scale the total active memory utilization rate, while c controls how predictable each event is from the past of X, with the time dependence statistically identifiable in the conditioning in both the dynamics $\mathbb{P}_X$ and $\mathbb{P}_X^0$. In this case as c increases, the more likely the refractory period is required to prevent a subsequent spike such that the processes characterized by $\mathbb{P}_X$ and $\mathbb{P}_X^0$ become more distinguishable leading to higher active memory utilization. Again, as a spiking process, the behavior of I_X is given by Eq. (61).

We note that an intrinsic feature of this example, after Y has been integrated out, has been its time inhomogeneity and subsequent nonstationary active memory utilization. This has crucially been dependent on the notions that (i) the stochastic behavior of the process can be time inhomogeneous (implemented here through a *deterministic* external process Y) and (ii) this time inhomogeneity can be statistically detected. This amounts to an ability to determine conditional dependence upon the *time* of evaluation such that, loosely, one can imagine that when conditioning on the sequence $x_{\mathcal{A}}$, one is always conditioning on both the history of X *and* the time, i.e., $\{x_A, \mathcal{A}\}$, and that one can, in theory, draw multiple realizations of the process starting from the same time origin. This notion is explored and formalized in Appendix B along with a discussion of what one should expect if such time dependence existed, but the ability to either condition on the time or, equivalently, draw multiple realizations starting at the same time origin was unavailable. In short we find that if this is the case, one always overestimates both the active memory utilization and transfer entropy rates. This is demonstrated for the specific model considered here in Appendix B 1 where we find that such an approach overestimates the active memory utilization rate by $(c/\Delta_y)\ln(\Delta_y/\Delta_x)$.

D. Full information dynamics description of a neural spiking model

Here we utilize a numerical model previously implemented in [22], but also calculate the active memory utilization alongside the transfer entropy, demonstrating their complementary

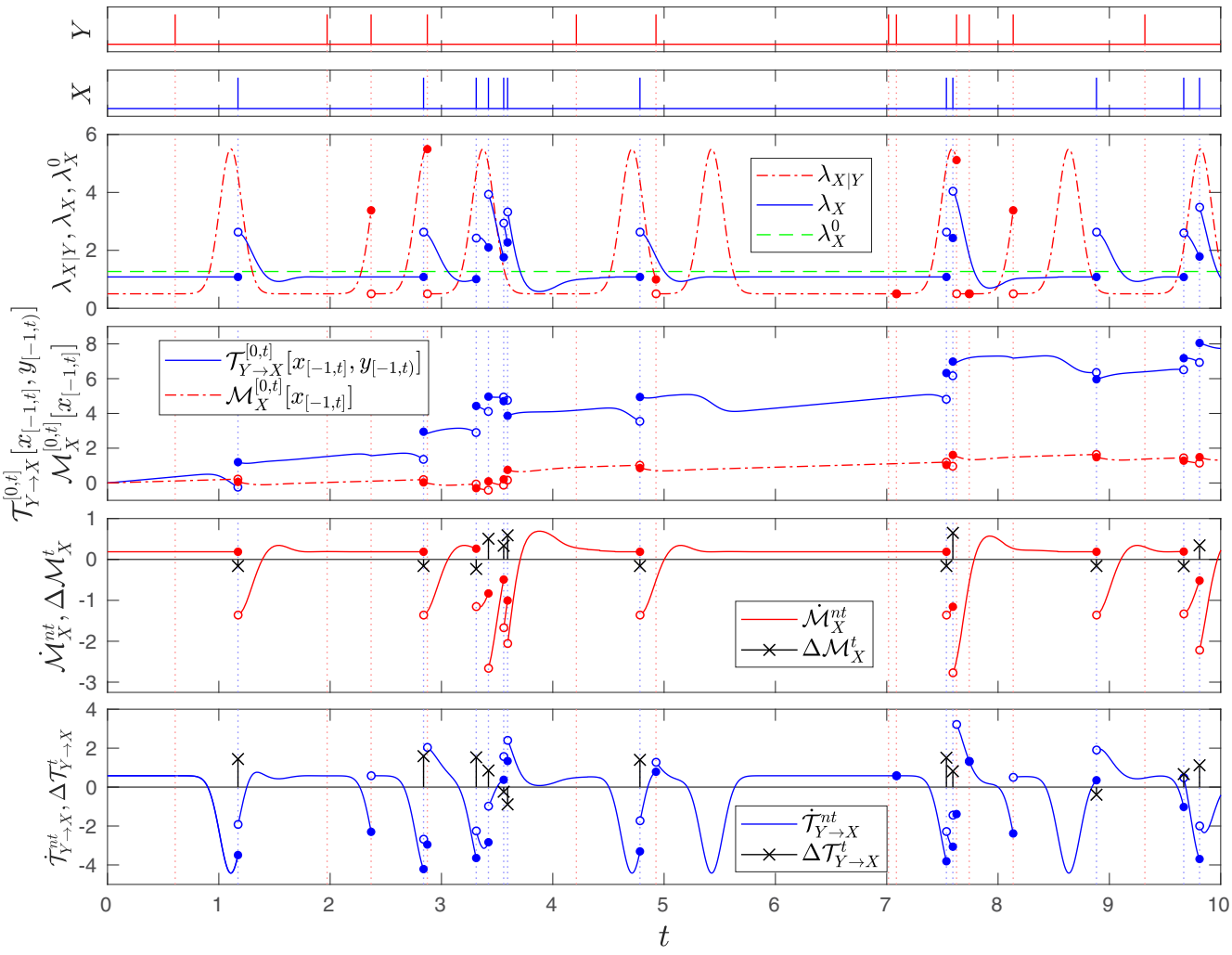

FIG. 1. Coupled spike trains generated using transition rates in Eq. (75) using $\lambda_Y = 1, \lambda_X^{\text{base}} = 0.5, m = 5, \sigma = 0.1, t_{\text{cut}} = 1$ along with generated and computed values of $\lambda_{X|Y}$, λ_X, and λ_X^0, resulting pathwise transfer entropy ($\mathcal{T}_{Y\to X}^{[0,t]}[x_{[-1,t]}, y_{[-1,t]}]$), pathwise active memory utilization ($\mathcal{M}_X^{[0,t]}[x_{[-1,t]}]$), and local contributions ($\Delta\mathcal{T}_{Y\to X}^t$, $\dot{\mathcal{T}}_{Y\to X}^{nt}$, $\Delta\mathcal{M}_X^t$, and $\dot{\mathcal{M}}_X^{nt}$). We set $\tau = -1$, matching the maximum historical dependence in λ_X and $\lambda_{X|Y}$, and a prior history of an absence of spikes in Y and X is assumed on the time interval $[-1,0)$. The lowermost and two uppermost panels reprinted and the third and fourth uppermost panels adapted from Ref. [22] with permission.

nature, illustrating the behavior of the pathwise quantities as a given pair of spike trains unfolds. In this model a spiking process, Y, follows a simple Poisson process characterized by a spike rate λ_Y (note therefore that Y possesses an active memory utilization rate of 0). Then the spiking process under consideration, X, spikes with a rate $\lambda_{X|Y}$ which depends upon the history of Y uniquely through the time *since* the last spike in Y, t^y:

$$\begin{aligned}
&\lambda_{X|Y}[x_{[\tau,t)}, y_{[\tau,t)}] \\
&\quad = \lambda_{X|Y}(t^y) \\
&\quad = \begin{cases} \lambda_X^{\text{base}}, & t^y \notin (0, t_{\text{cut}}], \\ \lambda_X^{\text{base}} + m \exp\left[-\frac{1}{2\sigma^2}\left(t^y - \frac{t_{\text{cut}}}{2}\right)^2\right], & t^y \notin (0, t_{\text{cut}}], \\ -m \exp\left[-\frac{1}{2\sigma^2}\left(\frac{t_{\text{cut}}}{2}\right)^2\right]. \end{cases} \qquad (75)
\end{aligned}$$

The detailed dependence is illustrated in Fig. 1, where the behavior of the pathwise active memory utilization and transfer entropy are contrasted on the interval $[0,t]$, $t \in [0,10]$, assuming a time origin $\tau = -1$ with no spikes in the history of X or Y in $[-1,0]$ (not shown) and that the system is in its stationary state. Explicitly, given the spike rate detailed in Eq. (75), we can identify (here in nats) the precise amount of information associated with memory utilization, $\mathcal{M}_X^{[0,t]}[x_{[-1,t]}]$, and signaling, $\mathcal{T}_{Y\to X}^{[0,t]}[x_{[-1,t]}, y_{[-1,t)}]$, on an arbitrary interval $[0,t]$ for the specific spiking behavior given in the first two subplots. The calculation of $\mathcal{M}_X^{[0,t]}[x_{[-1,t]}]$ relies on the ability to compute $\lambda_X[x_{[-1,t)}]$ and λ_X^0 at all times whereas the calculation of $\mathcal{T}_{Y\to X}^{[0,t]}[x_{[-1,t]}, y_{[-1,t)}]$ relies on the ability to compute $\lambda_{X|Y}[x_{[-1,t)}, y_{[-1,t)}]$ and $\lambda_X[x_{[-1,t)}]$ at all times. $\lambda_{X|Y}[x_{[-1,t)}, y_{[-1,t)}]$ is specified by Eq. (75), while λ_X^0, being a constant, since the process is stationary, is trivially determined by considering the mean number of spikes per unit time which is simply obtained by simulating the process and considering $\lambda_X^0 = \lim_{t\to\infty} N_x/t$, where N_x is the number of spikes in $[0,t]$, since the process is ergodic. Computation of $\lambda_X[x_{[0,t)}]$, however, is nontrivial, requiring a marginalization integration over $\lambda_{X|Y}[x_{[-1,t)}, y_{[-1,t)}]$ given the joint probabilistic behavior

of $\{x_{[-1,t)}, y_{[-1,t)}\}$. An algorithm to compute this was reported in [22] and utilized here.

Information associated with other intervals $[t',t]$, $0 < t' < t$, can be found by considering $\mathcal{M}_X^{[0,t]}[x_{[-1,t]}] - \mathcal{M}_X^{[0,t']}[x_{[-1,t']}]$ and $\mathcal{T}_{Y\to X}^{[0,t]}[x_{[-1,t]}, y_{[-1,t)}] - \mathcal{T}_{Y\to X}^{[0,t']}[x_{[-1,t']}, y_{[-1,t')}]$.

Looking at how these quantities evolve in time, we see discontinuous contributions to both the pathwise transfer entropy and active memory utilization when X spikes with continuous contributions in between them. The contributions to transfer entropy and active memory utilization are controlled by the relative sizes of $\lambda_{X|Y}$ and λ_X, and λ_X and λ_X^0, respectively. Positive discontinuous contributions in transfer entropy occur when $\lambda_{X|Y} > \lambda_X$ immediately preceding a spike in X while positive continuous contributions occur when $\lambda_{X|Y} < \lambda_X$. Similarly, positive discontinuous contributions to active memory utilization occur when $\lambda_X > \lambda_X^0$ immediately preceding a spike in X, while positive continuous contributions occur when $\lambda_X < \lambda_X^0$.

One distinct difference in behavior between $\mathcal{M}_X^{[0,t]}[x_{[-1,t]}]$ and $\mathcal{T}_{Y\to X}^{[0,t]}[x_{[-1,t]}, y_{[-1,t)}]$ can be observed in the evolution of the continuous contributions $\dot{\mathcal{M}}_X^{nt}$ and $\dot{\mathcal{T}}_{Y\to X}^{nt}$. $\dot{\mathcal{M}}_X^{nt}$ can only respond to changes in the history of X, while $\dot{\mathcal{T}}_{Y\to X}^{nt}$ can change in response to the history of both X and Y. Consequently, for this particular system, we only see discontinuities in $\dot{\mathcal{M}}_X^{nt}$ when X spikes; however we observe discontinuities in $\dot{\mathcal{T}}_{Y\to X}^{nt}$ when either X or Y spikes since a spike in X updates λ_X while a spike in Y updates $\lambda_{X|Y}$.

In the realization specified above the majority of the predictive capacity which can be associated with the full paths is being derived from the additional reduction in uncertainty Y provides through the transfer entropy, with a smaller residual predictive capacity being derived through the history of X through the active memory utilization.

Finally, we briefly mention the behavior of I_X, again given by Eq. (61) with limiting value $I_X = 0$. We can, however, for the particular numerical example in Fig. 1, state the numerical coefficient $c_{11} = 2\lambda_X^0$ by reporting the mean Markov spike rate $\lambda_X^0 \simeq 1.2697$.

VII. INFORMATION DYNAMICS IN GENERALIZED ORNSTEIN-UHLENBECK PROCESSES

One of the more striking corollaries of the preceding formalism is that purely Markov processes like the Ornstein-Uhlenbeck process in Eq. (14), while having nonzero, and indeed divergent, active information storage rates, have a vanishing memory utilization rate since, by definition in their construction, they only have dependence on their most recent state. In many respects this is appealing as the memory utilization rate then aligns very closely with the intuitive definition of a Markov process as being "memoryless."

However, if we couple multiple such Markov processes together, any individual process will no longer retain the Markov property due to the feedbacks between the processes. A simple example of such a model is that introduced in [41] consisting of two linearly coupled Ornstein-Uhlenbeck processes with correlated noise, viz.,

$$\begin{aligned} dx_t &= Ax_t dt + By_t dt + V_x dW_t^x, \\ dy_t &= Cx_t dt + Dy_t dt + V_y dW_t^y, \end{aligned} \tag{76}$$

where $\mathbb{E}[dW_t^x dW_{t'}^y] = \rho\delta(t-t')$ with $\rho \in [-1,1]$. The transfer entropy rate in the steady state of such a system is calculated in [41] and given by

$$\dot{T}_{Y\to X} = \frac{|D|}{2}\left[\sqrt{1 + \frac{BV_y}{DV_x}\left(\frac{BV_y}{DV_x} - 2\rho\right)} - \left(1 + \rho\frac{BV_y}{|D|V_x}\right)\right]. \tag{77}$$

With the transfer entropy and all subsequent quantities, the symmetry of the process allows us to identify all analogous quantities associated with Y ($\dot{T}_{X\to Y}$, $\dot{M}_Y$, I_Y) by making the substitutions $A \leftrightarrow D$, $B \leftrightarrow C$, and $V_x \leftrightarrow V_y$. To assess the full character of information processing in this system we wish to also consider $\dot{M}_x$ or, equivalently, $\dot{M}_x + \dot{T}_{Y\to X}$. The sum of these quantities is given by

$$\dot{M}_X + \dot{T}_{Y\to X} = \lim_{dt\to 0}\frac{1}{dt}\mathbb{E}\left[\ln\frac{p(x_{t+dt}|x_t, y_t)}{p(x_{t+dt}|x_t)}\right], \tag{78}$$

i.e., a Markov approximation to the transfer entropy given only the current values of the processes.

This reflects the property $\frac{d\mathbb{P}_{X|\{Y\}}}{d\mathbb{P}_X^0} = \frac{d\mathbb{P}_{X|\{Y\}}}{d\mathbb{P}_X}\frac{d\mathbb{P}_X}{d\mathbb{P}_X^0}$, since all the measures are equivalent, allowing us to write, in terms of pathwise quantities,

$$\begin{aligned} &\mathcal{M}_X^{[t_0,t]}[x_{[\tau,t]}] + \mathcal{T}_{Y\to X}^{[t_0,t]}[x_{[\tau,t]}, y_{[\tau,t)}] \\ &\quad = \mathcal{M}_X^{[t_0,t]}\left[x_{[t_0,t]}\right] + \mathcal{T}_{Y\to X}^{[t_0,t]}\left[x_{[t_0,t]}, y_{[t_0,t)}\right] \\ &\quad = \ln\frac{d\mathbb{P}_{X|\{Y\}}\left[x_{(t_0,t]}\middle|x_{t_0}, \{y_{[t_0,t)}\}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]}. \end{aligned} \tag{79}$$

Since both measures are Markovian and generate dynamics with the same sampling paths, with the same quadratic variation, etc., we may thus leverage the Cameron-Martin-Girsanov theorem in recognizing

$$\begin{aligned} &\frac{d\mathbb{P}_{X|\{Y\}}\left[x_{(t_0,t]}\middle|x_{t_0}, \{y_{[t_0,t)}\}\right]}{d\mathbb{P}_X^0\left[x_{(t_0,t]}\middle|x_{t_0}\right]} \\ &\quad = \exp\left[\frac{1}{2}\int_{t_0}^t f^2(x_{t'}, y_{t'})dt' + \int_{t_0}^t f(x_{t'}, y_{t'})dW_{t'}^x\right], \end{aligned} \tag{80}$$

such that

$$\begin{aligned} &d\mathcal{M}_X^{[t_0,t]}\left[x_{[t_0,t]}\right] + d\mathcal{T}_{Y\to X}^{[t_0,t]}\left[x_{[t_0,t]}, y_{[t_0,t)}\right] \\ &\quad = \tfrac{1}{2}f^2(x_t, y_t)dt + f(y_t, y_t)dW_t^x, \end{aligned} \tag{81}$$

where

$$f(x,y) = V_x^{-1}[Ax + By - \phi(x)], \tag{82}$$

and where $\phi(x)$ corresponds to the drift term in an effective dynamics which $\mathbb{P}_X^0$ describes, of the form

$$dx_t = \phi(x_t)dt + V_x dW_t^x. \tag{83}$$

Note, if we write $Z_t = \exp\{-\mathcal{M}_X^{[t_0,t]}[x_{[t_0,t]}] - \mathcal{T}_{Y\to X}^{[t_0,t]}[x_{[t_0,t]}, y_{[t_0,t)}]\}$, it immediately follows that Z_t is a Martingale, i.e.,

$$Z_t = 1 - \int_{t_0}^t Z_{t'} f(x_{t'}, y_{t'})dW_{t'}^x, \tag{84}$$

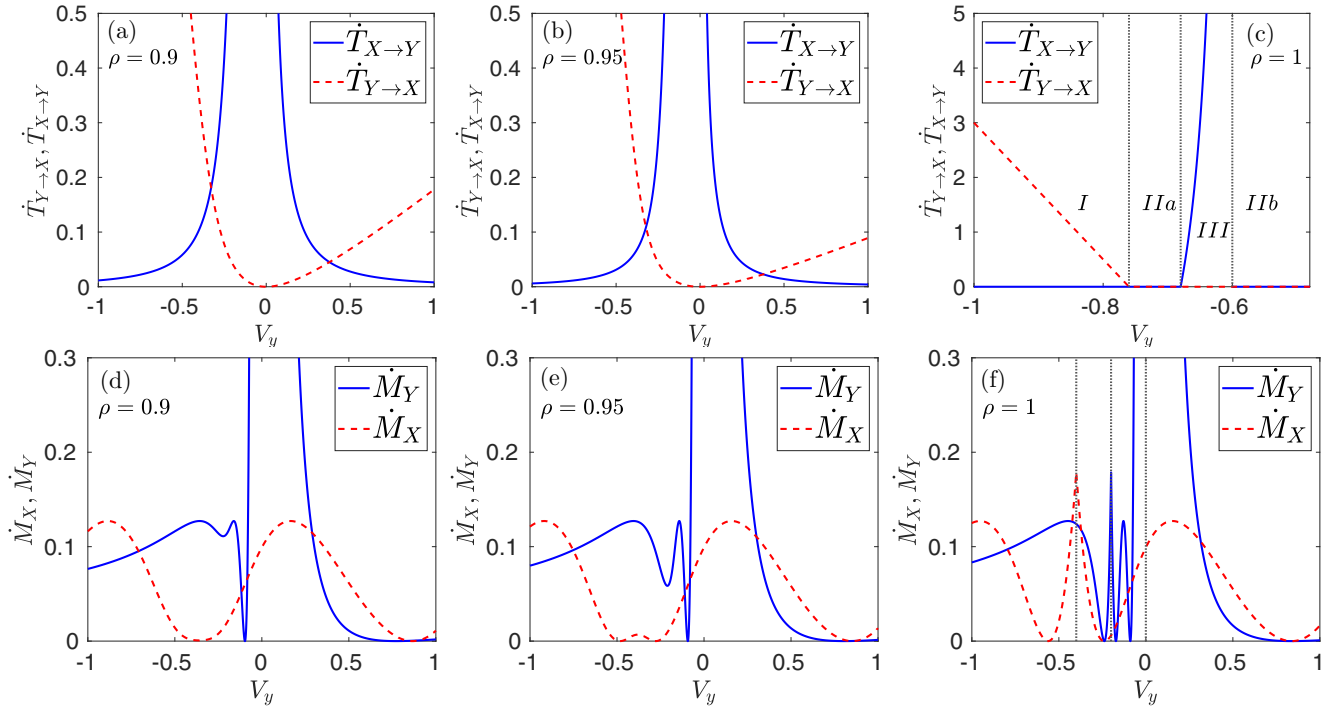

FIG. 2. Transfer entropy and active memory utilization rates for the correlated coupled Ornstein-Uhlenbeck process given by Eqs. (77) and (86) for varying noise correlation ρ and noise strength in the Y process V_y. We set $A = -5$, $B = 5$, $C = 1$, $D = -2$, $V_x = 1$. Panels (a), (b), and (c) show transfer entropy rates for noise correlation $\rho = 0.9$, $\rho = 0.95$, and $\rho = 1$, respectively. Panels (d), (e), and (f) show active memory utilization rates for noise correlation $\rho = 0.9$, $\rho = 0.95$, and $\rho = 1$, respectively.

implying $\mathbb{E}[Z_t] = 1$, not coincidentally mirroring the precise form of the so-called fluctuation theorems [42], due to their analogous construction based on RN derivatives.

The linear nature of the system dictates that the Markov marginal dynamics are also linear such that we have

$$\phi(x) = -\kappa_X^{\text{eff}} x,$$
$$\kappa_X^{\text{eff}} = \frac{(A+D)(BC-AD)V_x^2}{D(A+D)V_x^2 + B^2V_y^2 - BV_x(CV_x + 2\rho DV_y)}, \tag{85}$$

which can simply be read off the marginalized stationary distribution of the Fokker-Planck equation associated with the dynamics in Eq. (76) [41] or found by marginalizing the relevant short time transition probabilities or Green's functions and making appropriate manipulations using the stochastic calculus. From Eq. (80) it thus follows that

$$\begin{aligned}
&\dot{M}_X + \dot{T}_{Y\to X} \\
&= \tfrac{1}{2}\mathbb{E}[f^2(x,y)] = \left[4(A+D)V_x^2\right]^{-1} \\
&\quad \times \left[B^2V_y^2 + D(A+D)V_x^2 - BV_x(2D\rho V_y + CV_x)\right]^{-1} \\
&\quad \times B^2\{-B^2V_y^4 + 2B(D-A)\rho V_y^3V_x \\
&\quad - [(A+D)^2 - 2BC - 4AD\rho^2]V_x^2V_y^2 \\
&\quad - 2C(D-A)\rho V_yV_x^3 - C^2V_x^4\},
\end{aligned} \tag{86}$$

with the active memory utilization rate being the difference between Eqs. (86) and (77). The comparison of such terms yields rich structure even in such a simple system. Some of this structure is shown in Fig. 2 where the approach to complete correlation in the noise terms is shown for different noise strengths in Y. We note that both the transfer entropy and active memory utilization rate, $\dot{T}_{X\to Y}$ and $\dot{M}_Y$, diverge whenever $V_y \to 0$ since in this limit Y becomes a deterministic process such that the conditions required on RN derivatives for their existence are not met. In particular we point out the behavior when $\rho = 1$ where we observe different regimes in the character of information processing, marked in the top right panel. In regime I $\dot{T}_{Y\to X} > 0$ while $\dot{T}_{X\to Y} = 0$ and in regime III the opposite behavior is observed, $\dot{T}_{X\to Y} > 0$ while $\dot{T}_{Y\to X} = 0$. In the remaining regimes IIa and IIb all the transfer entropy rates are zero. The transitions between these regimes are marked with fine dashed lines. Interestingly, at the transitions separating regimes I and IIa and IIa and III we see a sharp, discontinuous, but not divergent, peak in the accompanying active memory utilization rates. The critical value of the noise strength in Y that separates regimes I and IIa is given by $V_y^{\text{crit}} = DV_x/B$, while the critical value of the noise strength in Y that separates regimes IIa and III is given by $V_y^{\text{crit}} = CV_x/A$.

Since the Markov marginal process is characterized by Eq. (83), specified with Eq. (85), i.e., a simple Ornstein-Uhlenbeck process as in Eq. (14), the contributions to the instantaneous predictive capacity, I_X, and components I_X^I and I_X^R, are equal to those in Eq. (30), but with $\kappa = \kappa_X^{\text{eff}}$ from Eq. (85).

VIII. DISCUSSION AND CONCLUSIONS

In this paper we have extended the approach elaborated in [22] for treating transfer entropy in continuous time to the broader framework of information dynamics. In doing so we have decomposed the active information storage into two distinct, positive, quantities called the active memory utilization and instantaneous predictive capacity. The former is complementary to the transfer entropy and inherits much of the behavior of transfer entropy in continuous time: there is a central cumulative quantity, the pathwise active memory utilization, associated with finite time intervals which possesses a mean rate at single instances in time. Individual behaviors, or events, are characterized by the pathwise active memory utilization, rather than a local rate, since the pathwise active memory utilization may be discontinuous. The latter quantity, the instantaneous predictive capacity, retains the predictive capacity of the process which does not assign finite contributions to individual path realizations, accounting for fundamental properties of the process such as the continuity properties of its sampling paths. Further, we have offered an asymptotic formalism for discussing this contribution, highlighting key differences in its structure for different processes. Since it accounts for intrinsic properties such as path continuity, which may lead to infinite predictive capacities, but has been derived in the context of separately maximizing all finite, pathwise, positive contributions, it is the minimal measure capable of offsetting similar effects in other measures of information processing, such as the excess entropy. Doing so reveals the maximum constituent component constructed from a cumulative pathwise quantity in the excess entropy is the so-called elusive information, which we have clarified should not be treated as a rate, but as an $O(1)$ quantity independently of the time basis.

We have then constructed such a formalism in the context of jump and neural spiking processes, complementing our previous work [22]. Using this we have demonstrated how to assess the complete information processing occurring in such a context comprising both memory and signaling. This has been illustrated in synthetic models of neural spiking demonstrating the qualitative behavior one should expect. Further, we have shown that the concepts offered here are well defined in other processes including coupled Ornstein-Uhlenbeck processes where we report interesting fine structure in the interplay between memory and signaling.

As with the transfer entropy, this work offers great promise particularly within the field of computational neuroscience where such a formalism lends itself to efficiently quantifying information processing in such settings. We finish by highlighting two particular consequences of our work in a neuroscience setting. First, as per estimation of the transfer entropy for neural spike trains [22], we anticipate that the most efficient estimation of $\dot{M}_X$ for such processes will emerge when utilizing an estimator utilizing interspike time intervals (which completely describe the process) as relevant continuous variables. Second, where active information storage is measured on discrete time samples of underlying continuous time processes (as is the case in neural imaging measurements), the active memory utilization is the only component that will approach a limiting value as the discrete time step approaches zero, and so may be the most appropriate quantity for investigation in such experiments.

ACKNOWLEDGMENTS

The authors thank Mikhail Prokopenko for useful discussions that contributed to this work. J.L. was supported through the Australian Research Council DECRA Grant No. DE160100630, and a Faculty of Engineering and IT Early Career Researcher and Newly Appointed Staff Development Scheme grant.

APPENDIX A: SELECTED DERIVATIONS OF RESULTS

1. Active information storage of the Ornstein-Uhlenbeck process

The process described by Eq. (14) permits a well known solution to its transition probability by consideration of the Green's function of the corresponding Fokker-Planck equation, due to the method of characteristics [29], given by

$$p_{OU}^{(\Delta t)}(x_{t+\Delta t}|x_t) = \sqrt{\frac{\kappa}{\pi\sigma^2(1-e^{-2\kappa\Delta t})}} \times \exp\left[-\frac{\kappa(x_{t+\Delta t} - x_t e^{-\kappa\Delta t})^2}{\sigma^2(1-e^{-2\kappa\Delta t})}\right], \quad \text{(A1)}$$

consistent with the stationary solution

$$p_{OU}(x_t) = \sqrt{\frac{\kappa}{\pi\sigma^2}}\exp\left[-\frac{\kappa x_t^2}{\sigma^2}\right] \quad \forall\, t. \quad \text{(A2)}$$

The active information storage is then simply given by the integral

$$A_X^{(\Delta t)} = \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dx\; p_{OU}^{(\Delta t)}(y|x) p_{OU}(x) \ln \frac{p_{OU}^{(\Delta t)}(y|x)}{p_{OU}(y)}, \quad \text{(A3)}$$

leading to the result in Eq. (15).

2. Instantaneous predictive capacity of discrete state master equations

Given the master equation on the discrete state space $\mathcal{X}$,

$$\dot{P}(x_i) = \sum_{x_j \neq x_i \in \mathcal{X}} W[x_i|x_j]P(x_j) - W[x_j|x_i]P(x_i), \quad \text{(A4)}$$

we can calculate the instantaneous predictive capacity over a short time Δt by considering up to 1 transition such that

$$I_X^{(\Delta t)} = \sum_{x_i \in \mathcal{X}} \sum_{x_j \neq x_i \in \mathcal{X}} P(x_i)\left[W[x_j|x_i]\Delta t \ln \frac{W[x_j|x_i]\Delta t}{P(x_j)} + (1 - \lambda[x_i]\Delta t)\ln\frac{(1-\lambda[x_i]\Delta t)}{P(x_i)}\right] + O(\Delta t^2), \quad \text{(A5)}$$

revealing the contributions

$$c_{00} = -\sum_{x_i \in \mathcal{X}} P(x_i)\ln P(x_i),$$

$$c_{10} = \sum_{x_i \in \mathcal{X}} P(x_i)\Big[\lambda[x_i][\ln P(x_i) - 1]$$

$$+\sum_{x_j\neq x_i\in\mathcal{X}} W[x_j|x_i]\ln\frac{W[x_j|x_i]}{P(x_j)}\Bigg]$$

$$=\sum_{x_i\in\mathcal{X}}\sum_{x_j\neq x_i\in\mathcal{X}} P(x_i)W[x_j|x_i]\left[\ln\frac{W[x_j|x_i]P(x_i)}{P(x_j)}-1\right],$$

$$c_{11}=\sum_{x_i\in\mathcal{X}}\sum_{x_j\neq x_i\in\mathcal{X}} P(x_i)W[x_j|x_i]. \tag{A6}$$

The instantaneous predictive capacity for the two species conversion process, $A \underset{k_+}{\overset{k_-}{\rightleftharpoons}} B$, utilized in Sec. IV B, can be derived from the dynamics underlying the master equation,

$$\dot{P}_A = k_- P_B - k_+ P_A, \quad \dot{P}_B = k_+ P_A - k_- P_B, \tag{A7}$$

i.e., $\mathcal{X}=\{A,B\}$, $W[A|B]=\lambda[B]=k_-$, $W[B|A]=\lambda[A]=k_+$ and with stationary solution $P_A = k_-/(k_-+k_+)$, $P_B = 1-P_A = k_+/(k_-+k_+)$ yielding the coefficients in Eq. (31).

APPENDIX B: CONDITIONAL AND NONSTATIONARY VARIANTS

Based on the formulation presented in the main text it is trivial to generalize, both the transfer entropy and active memory utilization, to the conditional case. In discrete time the conditional forms, conditioned upon some third variable Z, are given by

$$T_{Y\to X|Z}\equiv\mathbb{E}\left[\ln\frac{p(x_{n+1}|x_{\{0:n\}},y_{\{0:n\}},z_{\{0:n\}})}{p(x_{n+1}|x_{\{0:n\}},z_{\{0:n\}})}\right],\quad M_{X|Z}\equiv\mathbb{E}\left[\ln\frac{p(x_{n+1}|x_{\{0:n\}},z_{\{0:n\}})}{p(x_{n+1}|x_n,z_{\{0:n\}})}\right], \tag{B1}$$

which is straightforward to generalize to the continuous time case based on the generalization of the pathwise transfer entropy and pathwise active memory utilization to the *conditional pathwise transfer entropy* and *conditional pathwise active memory utilization*:

$$\mathcal{T}^{[t_0,t]}_{Y\to X|Z}[x_{[\tau,t]},y_{[\tau,t)},z_{[\tau,t)}]\equiv\ln\frac{d\mathbb{P}_{X|\{Y,Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{y_{[\tau,t)},z_{[\tau,t)}\}\right]}{d\mathbb{P}_{X|\{Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{z_{[\tau,t)}\}\right]},$$

$$\mathcal{M}^{[t_0,t]}_{X|Z}[x_{[\tau,t]},z_{[\tau,t)}]\equiv\ln\frac{d\mathbb{P}_{X|\{Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{z_{[\tau,t)}\}\right]}{d\mathbb{P}^0_{X|\{Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{z_{[\tau,t)}\}\right]}, \tag{B2}$$

where, analogously to the above, we may consider

$$\frac{d\mathbb{P}_{X|\{Y,Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{y_{[\tau,t)},z_{[\tau,t)}\}\right]}{d\mathbb{P}_{X|\{Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{z_{[\tau,t)}\}\right]}\sim\lim_{n\to\infty}\prod_{i=0}^{n}\frac{p(x_{i+1}|x_{\{-k:i\}},y_{\{-k:i\}},z_{\{-k:i\}})}{p(x_{i+1}|x_{\{-k:i\}},z_{\{-k:i\}})},$$

$$\frac{d\mathbb{P}_{X|\{Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{z_{[\tau,t)}\}\right]}{d\mathbb{P}^0_{X|\{Z\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{z_{[\tau,t)}\}\right]}\sim\lim_{n\to\infty}\prod_{i=0}^{n}\frac{p(x_{i+1}|x_{\{-k:i\}},z_{\{-k:i\}})}{p(x_{i+1}|x_i,z_{\{-k:i\}})}, \tag{B3}$$

where, again, $t_0=0$, $x_i\equiv x_{i\Delta t}$ and $\Delta t = t/(n+1) = -\tau/k$. Again we note the general construction of path probabilities with the $\{\}$ notation to mean $\mathbb{P}_{X|\{A\}}[x_{(t_0,t]}|x_{[\tau,t_0]},\{A_{[\tau,t)}\}]\sim\prod_{i=0}^{n}p(x_{i+1}|x_{\{-k:i\}},A_{\{-k:i\}})$ where A is some arbitrary extrinsic variable or variables in the form of a coincident time series.

This is, perhaps, not so illuminating in general; however it allows us to be precise when we discuss nonstationary transfer entropy and active memory utilization rates. To this end we make it clear that whenever such quantities are calculated, the time at which any transition probability is evaluated is also known such that one can construct them as conditional variants, conditioned on a third, deterministic, "variable," $\mathfrak{T}_t$, taking values t'_t equal to the time it is indexed by, i.e., $t'_t = t$. Thus, by conditioning on $\mathfrak{T}$ each relevant probability measure identifies any time dependence. As such, we explicitly take the following statements to be synonymous,

$$\mathcal{T}^{[t_0,t]}_{Y\to X}[x_{[\tau,t]},y_{[\tau,t)}]\equiv\mathcal{T}^{[t_0,t]}_{Y\to X|\mathfrak{T}}[x_{[\tau,t]},y_{[\tau,t)},t'_{[\tau,t)}],$$

$$\mathcal{M}^{[t_0,t]}_{X}[x_{[\tau,t]}]\equiv\mathcal{M}^{[t_0,t]}_{X|\mathfrak{T}}[x_{[\tau,t]},t'_{[\tau,t)}], \tag{B4}$$

such that it is only for brevity that $\mathfrak{T}$ is omitted in their formulation. We note that since $\mathfrak{T}$ merely represents the time index this alters some of the properties associated with the conditional probabilities used to construct the RN derivatives, namely that $t'_{\mathcal{A}}=\mathcal{A}$ and conditioning on $[\tau,t]$ is identical to conditioning on t. Accordingly we may write the RN derivatives underlying the quantities as

$$\exp\left[\mathcal{T}^{[t_0,t]}_{Y\to X|\mathfrak{T}}[x_{[\tau,t]},y_{[\tau,t)},t'_{[\tau,t)}]\right]=\frac{d\mathbb{P}_{X|\{Y,\mathfrak{T}\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{y_{[\tau,t)},[\tau,t)\}\right]}{d\mathbb{P}_{X|\mathfrak{T}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{[\tau,t)\}\right]}\sim\lim_{n\to\infty}\prod_{i=0}^{n}\frac{p(x_{i+1}|x_{\{-k:i\}},y_{\{-k:i\}},i\,\Delta t)}{p(x_{i+1}|x_{\{-k:i\}},i\,\Delta t)},$$

$$\exp\left[\mathcal{M}^{[t_0,t]}_{X|\mathfrak{T}}[x_{[\tau,t]},t'_{[\tau,t)}]\right]=\frac{d\mathbb{P}_{X|\{\mathfrak{T}\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{[\tau,t)\}\right]}{d\mathbb{P}^0_{X|\{\mathfrak{T}\}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{[\tau,t)\}\right]}\sim\lim_{n\to\infty}\prod_{i=0}^{n}\frac{p(x_{i+1}|x_{\{-k:i\}},i\,\Delta t)}{p(x_{i+1}|x_i,i\,\Delta t)}. \tag{B5}$$

It is common, however, to assume stationarity in a process when these quantities are computed empirically. This amounts to using the alternative measures $\mathbb{P}^{\mathrm{st}}_{X|\{Y\}}$, $\mathbb{P}^{\mathrm{st}}_{X}$, and $\mathbb{P}^{0,\mathrm{st}}_{X}$, constructed by averaging the probabilistic behavior experienced at all observed

times, assuming equal *a priori* probabilities with respect to any instance in time. This means that the estimated quantities converge to different "stationary" quantities denoted $\mathcal{T}_{Y\to X}^{\mathrm{st},[t_0,t]}[x_{[\tau,t]},y_{[\tau,t]}]$ and $\mathcal{M}_X^{\mathrm{st},[t_0,t]}[x_{[\tau,t]}]$, and, $\dot{T}_{Y\to X}^{\mathrm{st}}$ and $\dot{M}_X^{\mathrm{st}}$ defined through

$$\begin{aligned}\mathcal{T}_{Y\to X}^{\mathrm{st},[t_0,t]}[x_{[\tau,t]},y_{[\tau,t]}] &\equiv \ln\frac{d\mathbb{P}_{X|\{Y\}}^{\mathrm{st}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]},\{y_{[\tau,t)}\}\right]}{d\mathbb{P}_X^{\mathrm{st}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]} \sim \lim_{n\to\infty}\ln\left[\prod_{i=0}^{n}\frac{p^{\mathrm{st}}(x_{i+1}|x_{\{-k:i\}},y_{\{-k:i\}})}{p^{\mathrm{st}}(x_{i+1}|x_{\{-k:i\}})}\right]\\ &= \lim_{n\to\infty}\ln\left[\prod_{i=0}^{n}\lim_{n\to\infty}\frac{\sum_{j=-k}^{n}p(x_{i+1}|x_{\{-k:i\}},y_{\{-k:i\}},j\Delta t)}{\sum_{j=-k}^{n}p(x_{i+1}|x_{\{-k:i\}},j\Delta t)}\right]\end{aligned} \tag{B6}$$

and

$$\begin{aligned}\mathcal{M}_X^{\mathrm{st},[t_0,t]}[x_{[\tau,t]}] &\equiv \ln\frac{d\mathbb{P}_X^{\mathrm{st}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]}{d\mathbb{P}_X^{0,\mathrm{st}}\left[x_{(t_0,t]}\middle|x_{[\tau,t_0]}\right]} \sim \lim_{n\to\infty}\ln\left[\prod_{i=0}^{n}\frac{p^{\mathrm{st}}(x_{i+1}|x_{\{-k:i\}})}{p^{\mathrm{st}}(x_{i+1}|x_i)}\right]\\ &= \lim_{n\to\infty}\ln\left[\prod_{i=0}^{n}\lim_{n\to\infty}\frac{\sum_{j=-k}^{n}p(x_{i+1}|x_{\{-k:i\}},j\Delta t)}{\sum_{j=-k}^{n}p(x_{i+1}|x_i,j\Delta t)}\right].\end{aligned} \tag{B7}$$

That is, we understand that the empirically computed probabilities, assuming stationarity, will approximate

$$p^{\mathrm{st}}(x_{i+1}|x_{\{-k:i\}}) = \lim_{n\to\infty}\frac{1}{n+k+1}\sum_{j=-k}^{n}p(x_{i+1}|x_{\{-k:i\}},t=j\Delta t). \tag{B8}$$

Clearly, however, if only one or a limited number of samples are available, this approximation cannot be expected to be accurate, in the general case, however long the samples, unless, for example, the underlying time variation in p is periodic or, if controlled by some hidden variable, that variable evolves ergodically. As such, only when a process is stationary do we explicitly have $\mathcal{T}_{Y\to X}^{\mathrm{st},[t_0,t]}=\mathcal{T}_{Y\to X}^{[t_0,t]}$ and $\dot{T}_{Y\to X}^{\mathrm{st}}=\dot{T}_{Y\to X}$, and, $\mathcal{M}_X^{\mathrm{st},[t_0,t]}=\mathcal{M}_X^{[t_0,t]}$ and $\dot{M}_X^{\mathrm{st}}=\dot{M}_X$. Moreover, we have formulated the difference between these quantities in terms of a marginalization over an implied process, $\mathfrak{T}$. Consequently, treating this process like any other, we can identify the difference between the "stationary" and nonstationary quantities as

$$\dot{T}_{Y\to X}^{\mathrm{st}} - \mathring{T}_{Y\to X} = \dot{T}_{\mathfrak{T}\to X} - \dot{T}_{\mathfrak{T}\to X|Y} \tag{B9}$$

and

$$\dot{M}_X^{\mathrm{st}} - \mathring{M}_X = \dot{T}_{\mathfrak{T}\to X}^{(0)} - \dot{T}_{\mathfrak{T}\to X} \tag{B10}$$

where, analogously,

$$\begin{aligned}\dot{T}_{\mathfrak{T}\to X} &= \lim_{t\to\infty}\frac{1}{t-t_0}\int_{t_0}^{t}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\left[\ln\frac{p(x_{t'+\Delta t}|x_{[\tau,t')},t')}{p^{\mathrm{st}}(x_{t'+\Delta t}|x_{[\tau,t')})}\right]dt'\\ &= \lim_{t\to\infty}\frac{1}{t-t_0}\int_{t_0}^{t}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\left[\ln\frac{p(x_{t'+\Delta t}|x_{[\tau,t')},t')}{\lim_{t\to\infty}(t-t_0)^{-1}\int_{t_0}^{t}p(x_{t'+\Delta t}|x_{[\tau,t')},t'')dt''}\right]dt',\\ \dot{T}_{\mathfrak{T}\to X}^{(0)} &= \lim_{t\to\infty}\frac{1}{t-t_0}\int_{t_0}^{t}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\left[\ln\frac{p(x_{t'+\Delta t}|x_{t'},t')}{p^{\mathrm{st}}(x_{t'+\Delta t}|x_{t'})}\right]dt'\\ &= \lim_{t\to\infty}\frac{1}{t-t_0}\int_{t_0}^{t}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\left[\ln\frac{p(x_{t'+\Delta t}|x_{t'},t')}{\lim_{t\to\infty}(t-t_0)^{-1}\int_{t_0}^{t}p(x_{t'+\Delta t}|x_{t'},t'')dt''}\right]dt'.\end{aligned} \tag{B11}$$

In particular, if $\dot{T}_{\mathfrak{T}\to X}=0$, such that an infinite history length in X allows us predict equally well with or without a time index, we risk overestimation of the active memory utilization since $\dot{M}_X^{\mathrm{st}}-\mathring{M}_X=\dot{T}_{\mathfrak{T}\to X}^{(0)}\geqslant 0$.

It is important to note that it may be challenging, at least empirically, to identify such a distinction between stationary and nonstationary formulations, as it requires us to be able to probe the statistics of the process at a given time t. The ability to achieve this requires the ability (i) to hypothetically draw multiple samples or realizations from the generating process starting at the same time origin and (ii) to implicitly allow access to the time of evaluation when considering any transition probability. Indeed this may be deemed completely impossible in practice leaving such a question fundamentally ambiguous.

1. Stationary active memory utilization calculation for the model utilized in Sec. VI C 2

We can illustrate the above distinctions and possible ambiguity by calculating $\dot{M}_X^{\mathrm{st}}$ for the process described in Sec. VI C 2 and discuss when and how the distinction between the two calculations would be the same, different, or unknowable. In this system, nonstationary behavior is introduced by means of a deterministic variable Y upon which the behavior in X depends. The system can be identified as nonstationary by appealing to an ensemble of realizations at any given time t. We can, however, imagine

that the time indexing of such an ensemble is either not known or not knowable, equivalent to the assumption that the system is stationary if the spike rates were to be constructed empirically from data.

To calculate $\dot{M}_X^{\mathrm{st}}$, therefore, requires calculation of two analogous spike rates λ_X^{st} and $\lambda_X^{\mathrm{st},0}$ where only the sequences of the relevant path histories are known, and not the time at which they are being evaluated. Calculation of $\lambda_X^{\mathrm{st},0}$ is straightforward and amounts to the aggregated mean spike rate in X, which can either just be asserted by recognizing that there are, on average, $c \times (T/\Delta_y)$ spikes in an interval T in the $T \to \infty$ limit or by writing

$$\lambda_X^{\mathrm{st},0} = \lim_{t-t_0\to\infty} \frac{1}{t-t_0}\int_{t_0}^{t} \lambda_X^0(t')dt = \frac{1}{\Delta_y}\int_{n\Delta_y}^{(n+1)\Delta_y} \lambda_X^0(t')dt' = \frac{c}{\Delta_y}. \tag{B12}$$

On the other hand, λ_X^{st} depends on how much path history is available to condition upon. We will take the limit of a time origin $\tau = t_0 \to -\infty$, both for simplicity and because such a condition will dominate in the case $t \to \infty$ when considering long time, steady state behavior. We find such behavior in several steps. First, noting again the shorthand of Eq. (69), we construct

$$\lambda_X^{\mathrm{st}}\left[x_{[t_0,t)}\right] = \frac{1}{p^{\mathrm{st}}\left[x_{[t_0,t)}\right]}\int dy_{[t_0,t)}\lambda_{X|Y}^{\mathrm{st}}\left[x_{[t_0,t)}\right],y_{[t_0,t)}]p^{\mathrm{st}}\left[x_{[t_0,t)},y_{[t_0,t)}\right] = \int dy_{[t_0,t)}\lambda_{X|Y}\left[x_{[t_0,t)},y_{[t_0,t)}\right]\frac{p^{\mathrm{st}}\left[x_{[t_0,t)}\middle|y_{[t_0,t)}\right]p^{\mathrm{st}}\left[y_{[t_0,t)}\right]}{p^{\mathrm{st}}\left[x_{[t_0,t)}\right]}, \tag{B13}$$

where we have recognized $\lambda_{X|Y}^{\mathrm{st}}[x_{[t_0,t)},y_{[t_0,t)}] = \lambda_{X|Y}[x_{[t_0,t)},y_{[t_0,t)}]$ due to the ability to write $\lambda_{X|Y}[x_{[t_0,t)},y_{[t_0,t)}]$ independently of t as per Sec. VI C 2. Next we consider the form of $p^{\mathrm{st}}[y_{[t_0,t)}]$. Recall Y always realizes $y^*_{[t_0,t)} = \{\ldots,-\Delta_y,0,\Delta_y,2\Delta_y,\ldots\}$ such that $p[y_{[t_0,t)}] = \delta(y_{[t_0,t)} - y^*_{[t_0,t)})$. $p^{\mathrm{st}}[y_{[t_0,t)}]$ is then the probability of the sequence of $y^*_{[t_0,t)}$ disassociated with its timing such that we do not know if we are considering the probability of the *sequence* $y^*_{[t_0,t)}$ at time t or any other time. Consequently, $p^{\mathrm{st}}[y_{[t_0,t)}]$ assigns probability to every path $y_{[t_0,t)}$ that consists of spikes at precise intervals of Δ_y, but is shifted by an arbitrary phase factor $\phi \in [0,\Delta_y)$, i.e., $y_{[t_0,t)} = y^*_{[t_0+\phi,t+\phi)}$, according to the distribution $p_\Phi(\phi)$ (which is necessarily flat given equal *a priori* probability at all times, though left general for completeness and later discussion). That is, all variation in $y_{[t_0,t)}$ is through the single parameter ϕ. Consequently we have $p^{\mathrm{st}}[x_{[t_0,t)},y_{[t_0,t)}] = p^{\mathrm{st}}[x_{[t_0,t)}|\phi]p_\Phi(\phi)$ and thus

$$\lambda_X^{\mathrm{st}}\left[x_{[t_0,t)}\right] = \int_0^{\Delta_y} d\phi\ \lambda_{X|Y}\left[x_{[t_0,t)},\phi\right]\frac{p^{\mathrm{st}}\left[x_{[t_0,t)}\middle|\phi\right]p_\Phi(\phi)}{p^{\mathrm{st}}\left[x_{[t_0,t)}\right]}. \tag{B14}$$

At this point we make the critical claim

$$\lim_{t_0\to-\infty} p^{\mathrm{st}}\left[x_{[t_0,t)}\middle|\phi\right]p_\Phi(\phi) = \lim_{t_0\to-\infty} p^{\mathrm{st}}\left[x_{[t_0,t)}\right]\delta(\phi), \tag{B15}$$

which is to say, given a long enough past sequence $x_{[t_0,t)}$ there is only a single compatible value of ϕ. Since the process is actually generating $y^*_{[t_0,t)}$ then this value must be $\phi = 0$. This can be seen by posing the (inverse) question: given an arbitrarily long sequence $x_{[t_0,t)}$ (along with knowledge of its time indexing) does there exist a way of uniquely determining ϕ? The answer is yes with it being obtained by searching the past sequence of X for the minimum interspike interval. In the infinite limit this is the smallest *possible* interspike interval. This occurs when the first of such spikes coincides with the very end of a possible spiking window following a spike in Y, with timing $t = n\Delta_y + \Delta_x + \phi$, and the subsequent spike occurring at the very beginning of the next possible spiking window following a spike in Y, with timing $t = (n+1)\Delta_y + \phi$ (with $n \in \mathbb{Z}$). Consequently,

$$\begin{aligned}\lim_{t_0\to-\infty}\lambda_X^{\mathrm{st}}\left[x_{[t_0,t)}\right] &= \lim_{t_0\to-\infty}\int_0^{\Delta_y} d\phi\,\lambda_{X|Y}\left[x_{[t_0,t]},\phi\right]\frac{p^{\mathrm{st}}\left[x_{[t_0,t]}\right]}{p^{\mathrm{st}}\left[x_{[t_0,t]}\right]}\delta(\phi) = \lim_{t_0\to-\infty}\lambda_{X|Y}\left[x_{[t_0,t]},\phi=0\right]\\ &= \lambda_{X|Y}\left[x_{[t_0,t]},y^*_{[\tau,t)}\right] = \lambda_{X|Y}(t_x,t_y) = \lambda_X(t_x,t);\end{aligned} \tag{B16}$$

i.e., all the predictive capability of $\lambda_{X|Y}$ (and thus λ_X) can be gleaned from the history of X even without knowledge of the current time.

We then use these two stationary transition rates to construct the average associated with the calculation of $\dot{M}_X^{\mathrm{st}}$. However, since $\lambda_X^{\mathrm{st},0}$ is just a constant and $\lim_{t_0\to-\infty}\lambda_X^{\mathrm{st}}[x_{[t_0,t)}] = \lambda_X(t_x,t)$, this allows us to greatly simplify the implied average over all infinitely long paths $x_{[t_0,t)}$ by replacing $\lambda_X^{\mathrm{st}}[x_{[t_0,t)}]$ with $\lambda_X(t_x,t)$ and integrating over a single period $[n\Delta_y,(n+1)\Delta_y]$. As such we find

$$\begin{aligned}\dot{M}_X^{\mathrm{st}} &= \lim_{t_0\to-\infty}\mathbb{E}\left[\left(1-\delta_{x_t,x_t^-}\right)\ln\frac{\lambda_X^{\mathrm{st}}\left[x_{[t_0,t)}\right]}{\lambda_X^{\mathrm{st},0}}\right] = \lim_{t_0\to-\infty}\int dx_{[t_0,t)}p^{\mathrm{st}}\left[x_{[t_0,t)}\right]\lambda_X^{\mathrm{st}}\left[x_{[t_0,t)}\right]]\ln\frac{\lambda_X^{\mathrm{st}}\left[x_{[t_0,t)}\right]}{\lambda_X^{\mathrm{st},0}}\\ &= \frac{1}{\Delta_y}\int_{n\Delta_y}^{(n+1)\Delta_y}\mathbb{E}\left[\left(1-\delta_{x_t,x_t^-}\right)\ln\frac{\lambda_X(t_x,t)}{\lambda_X^{\mathrm{st},0}}\right]dt = \frac{1}{\Delta_y}\int_{n\Delta_y}^{(n+1)\Delta_y}\\ &\quad\times\exp\left[-\int_{n\Delta_y}^{t}\lambda_X(t_x\leqslant t-\Delta_x,t')dt'\right]\lambda_X(t_x\leqslant t-\Delta_x,t)\ln\frac{\Delta_y\lambda_X(t_x\leqslant t-\Delta_x,t)}{c}dt\end{aligned}$$

$$= \frac{1}{\Delta_y}\int_{n\Delta_y}^{n\Delta_y+\Delta_x} \frac{\Delta_x - c(t-n\Delta_y)}{\Delta_x}\lambda_X(t_x \leqslant t-\Delta_x,t)\ln\frac{\Delta_y\lambda_X(t_x\leqslant t-\Delta_x,t)}{c}dt$$
$$= (\Delta_y)^{-1}\{(1-c)\ln(1-c) + c[1+\ln(\Delta_y/\Delta_x)]\}. \tag{B17}$$

Here we see an additional term as compared to the nonstationary result such that $\dot{M}_X^{\rm st} = \mathring{M}_X + (c/\Delta_y)\ln(\Delta_y/\Delta_x)$. This extra contribution over $\dot{M}_X$ arises from the ability of the full dynamics in X to distinguish whether the system was in the spiking window, $[n\Delta_y, n\Delta_y+\Delta_x]$, or not over the time homogeneous Markov marginalization process, characterized by $\lambda_X^{{\rm st},0}$, which cannot detect either the refractory period or the existence of this window. Since $\lim_{t_0\to-\infty}\lambda_X[x_{[t_0,t)}] = \lambda(t_x,t)$, by definition, $\dot{T}_{\mathfrak{T}\to X} = 0$, thus illustrating the specific case $\dot{M}_X^{\rm st} \geqslant \mathring{M}_X$ emerging from Eq. (B10) as claimed earlier in this Appendix.

As we have seen, the process described in Sec. VI C 2 leads to a disparity between $\dot{M}_X^{\rm st}$ and $\mathring{M}_X$ due to the time inhomogeneous spike rate in X, corresponding, in the framework described here, to $p_\Phi(\phi) = \delta(\phi)$. However, we can consider simple alterations to this process which change this property. Importantly, Eq. (B15) holds for any $p_\Phi(\phi)$ and thus so does the final relation in Eq. (B16) with the only exception being that ϕ need not equal 0, but corresponds to whatever value of ϕ is drawn from $p_\Phi(\phi)$. As such if we consider a process identical to that in Sec. VI C 2 but where $y_{[t_0,t)}$ is generated such that it equals $y^*_{[\tau+\phi,t+\phi)}$ with probability density $p_\Phi(\phi)$, it will have $\dot{M}_X^{\rm st}$ equal to that in Eq. (B17) independently of $p_\Phi(\phi)$. Consequently, if we choose $p_\Phi(\phi) = \Delta_y^{-1}$, such that Y and thus X are both stationary, we will have $\mathring{M}_X = \dot{M}_X = \dot{M}_X^{\rm st}$, again with $\dot{M}_X^{\rm st}$ given by Eq. (B17).

Indeed we can give an expression for $\dot{M}_X$ for arbitrary $p_\Phi(\phi)$ other than the $p_\Phi(\phi) = \Delta_y^{-1}$ and $p_\Phi(\phi) = \delta(\phi)$ we have already considered. To do this we first extend the domain of $p_\Phi(\phi)$ to $\phi \in \mathbb{R}$ such that it is periodic, i.e., $p_\Phi(\phi + n\Delta_y)$ with $n \in \mathbb{Z}$, but retaining normalization on $[0,\Delta_y)$ [i.e., $\int_0^{\Delta_y} p_\Phi(\phi')d\phi' = \int_\phi^{\phi+\Delta_y} p_\Phi(\phi')d\phi' = 1$]. We can then write an expression for $\lambda_X^0(t)$ as

$$\lambda_X^0(t) = \frac{c}{\Delta_x}\int_{t-\Delta_x}^{t} p_\Phi(\phi)d\phi, \tag{B18}$$

thus expressing the difference $\dot{M}_X^{\rm st} - \mathring{M}_X$ as

$$\dot{M}_X^{\rm st} - \mathring{M}_X = \int_0^{\Delta_y}\left[\frac{1}{\Delta_y}\int_\phi^{\phi+\Delta_y} e^{-\int_\phi^t \lambda_X(t_x\leqslant t-\Delta_x,t')dt'}\lambda_X(t_x\leqslant t-\Delta_x,t)\ln\frac{\lambda_X^0(t)}{\lambda_X^{{\rm st},0}}dt\right]p_\Phi(\phi)d\phi \geqslant 0, \tag{B19}$$

where we have noted a lower bound due to its expression as a KL divergence, or more particularly, its expression as $\dot{T}^{(0)}_{\mathfrak{T}\to X}$, due to the fact $\dot{T}_{\mathfrak{T}\to X} = 0$ as per Eq. (B10).

Continuing, we have

$$\dot{M}_X^{\rm st} - \mathring{M}_X = \int_0^{\Delta_y}\left[\frac{1}{\Delta_y}\int_\phi^{\phi+\Delta_x}\frac{c}{\Delta_x}\ln\frac{\lambda_X^0(t)}{\lambda_X^{{\rm st},0}}dt\right]p_\Phi(\phi)d\phi = \frac{1}{\Delta_y}\int_0^{\Delta_y} f(\phi)p_\Phi(\phi)d\phi,$$
$$f(\phi) = \int_\phi^{\phi+\Delta_x}\frac{c}{\Delta_x}\ln\left[\frac{\Delta_y}{\Delta_x}\int_{t-\Delta_x}^{t} p_\Phi(\phi')d\phi'\right]dt, \tag{B20}$$

such that

$$\dot{M}_X^{\rm st} - \mathring{M}_X = \frac{c}{\Delta_y}\ln\frac{\Delta_y}{\Delta_x} + \xi, \tag{B21}$$

with

$$\xi = \frac{c}{\Delta_y\Delta_x}\int_0^{\Delta_y} p_\Phi(\phi)\int_\phi^{\phi+\Delta_x}\ln\left[\int_{t-\Delta_x}^{t} p_\Phi(\phi')d\phi'\right]dtd\phi. \tag{B22}$$

The contents of the logarithm in ξ always lie in [0,1], due to the assertion $2\Delta_x < \Delta_y$ in the construction of the model, and so we have $\xi \leqslant 0$. ξ takes a maximum value 0 when $p_\Phi(\phi) = \delta(\phi - a)$, $a \in [0,\Delta_y)$, with $a = 0$ corresponding to the usage in Sec. VI C 2. Due to Eq. (B19) being a KL divergence, it therefore takes a minimum value $-\frac{c}{\Delta_y}\ln\frac{\Delta_y}{\Delta_x}$ corresponding to $p_\Phi(\phi) = \Delta_y^{-1}$, i.e., when the process is stationary, such that

$$-\frac{c}{\Delta_y}\ln\frac{\Delta_y}{\Delta_x} \leqslant \xi \leqslant 0, \tag{B23}$$

and in turn

$$0 \leqslant \dot{M}_X^{\rm st} - \mathring{M}_X \leqslant \frac{c}{\Delta_y}\ln\frac{\Delta_y}{\Delta_x}. \tag{B24}$$

It is worth pointing out that such a process, despite being stationary when $p_\Phi(\phi) = \Delta_y^{-1}$, would not be ergodic for any choice of $p_\Phi(\phi)$ since once ϕ is drawn from the distribution, the process deterministically spikes with period Δ_y indefinitely.

Finally, if only one sample, $\{x_{[\tau,t)}, y_{[\tau,t)}\}$, is drawn, however long, from which empirical estimates are to be formed, then there is fundamental ambiguity as to the statistical nature of Y and thus how large the overestimation, or underestimation, of the active memory utilization rate, $\dot{M}_X^{\text{st}} - \mathring{M}_X$, is. *If* it can be asserted that Y is indeed drawn from a distribution $p_\Phi(\phi)$, only then may we state that it is an overestimation that lies in $[0, \frac{c}{\Delta_y} \ln \frac{\Delta_y}{\Delta_x}]$ as per the above.

[1] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Phys. Rev. E **77**, 026110 (2008).

[2] J. T. Lizier, in *Directed Information Measures in Neuroscience*, Understanding Complex Systems, edited by M. Wibral, R. Vicente, and J. T. Lizier (Springer, Berlin, 2014), pp. 161–193.

[3] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Chaos **20**, 037109 (2010).

[4] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Information Sciences **208**, 39 (2012).

[5] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[6] J. T. Lizier, *The Local Information Dynamics of Distributed Computation in Complex Systems* (Springer, Berlin, 2013).

[7] J. T. Lizier, S. Pritam, and M. Prokopenko, Artificial Life **17**, 293 (2011).

[8] P. L. Williams and R. D. Beer, in *From Animals to Animats 11*, Lecture Notes in Computer Science, Vol. 6226, edited by S. Doncieux, B. Girard, A. Guillot, J. Hallam, J.-A. Meyer, and J.-B. Mouret (Springer, Berlin, 2010), pp. 38–49.

[9] S. Dasgupta, F. Wörgötter, and P. Manoonpong, Evolving Systems **4**, 235 (2013).

[10] S. I. Walker, H. Kim, and P. C. W. Davies, Philos. Trans. R. Soc. A **374**, 20150057 (2016).

[11] T. Tomaru, H. Murakami, T. Niizato, Y. Nishiyama, K. Sonoda, T. Moriyama, and Y.-P. Gunji, Artificial Life and Robotics **21**, 177 (2016).

[12] X. R. Wang, J. M. Miller, J. T. Lizier, M. Prokopenko, and L. F. Rossi, PLoS One **7**, e40084 (2012).

[13] J. Garland, R. G. James, and E. Bradley, Phys. Rev. E **93**, 022221 (2016).

[14] L. Faes and A. Porta, in *Directed Information Measures in Neuroscience*, Understanding Complex Systems, edited by M. Wibral, R. Vicente, and J. T. Lizier (Springer, Berlin, 2014), pp. 61–86.

[15] N. Timme, S. Ito, M. Myroshnychenko, F.-C. Yeh, E. Hiolski, P. Hottowy, and J. M. Beggs, PLoS One **9**, e115764 (2014).

[16] S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs, PLoS One **6**, e27431 (2011).

[17] L. Faes, D. Kugiumtzis, G. Nollo, F. Jurysta, and D. Marinazzo, Phys. Rev. E **91**, 032904 (2015).

[18] M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser, Prog. Biophys. Mol. Biol. **105**, 80 (2011).

[19] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, J. Comput. Neurosci. **30**, 45 (2011).

[20] A. Brodski-Guerniero, G.-F. Paasch, P. Wollstadt, I. Özdemir, J. T. Lizier, and M. Wibral, J. Neurosci. **37**, 8273 (2017).

[21] C. Gómez, J. T. Lizier, M. Schaum, P. Wollstadt, C. Grützner, P. Uhlhaas, C. M. Freitag, S. Schlitt, S. Bölte, R. Hornero, and M. Wibral, Frontiers in Neuroinformatics **8**, 9 (2014).

[22] R. E. Spinney, M. Prokopenko, and J. T. Lizier, Phys. Rev. E **95**, 032319 (2017).

[23] L. Barnett and A. K. Seth, J. Neurosci. Methods **275**, 93 (2017).

[24] D. P. Feldman, C. S. McTague, and J. P. Crutchfield, Chaos **18**, 043106 (2008).

[25] A. Kaiser and T. Schreiber, Phys. D (Amsterdam, Neth.) **166**, 43 (2002).

[26] L. Barnett and T. Bossomaier, Phys. Rev. Lett. **109**, 138105 (2012).

[27] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy: Information Flow in Complex Systems* (Springer International Publishing, Cham, Switzerland, 2016).

[28] M. Wibral, J. T. Lizier, S. Vögler, V. Priesemann, and R. Galuske, Front. Neuroinform. **8**, 1 (2014).

[29] H. Risken and T. Frank, *The Fokker-Planck Equation: Methods of Solution and Applications*, 2nd ed., Springer Series in Synergetics (Springer-Verlag, Berlin, 1996).

[30] S. Marzen and J. P. Crutchfield, Entropy **16**, 4713 (2014).

[31] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals* (Dover Publications, Inc., New York, 1986).

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Inc., 2005).

[33] J. P. Crutchfield and D. P. Feldman, Chaos **13**, 25 (2003).

[34] W. Bialek, I. Nemenman, and N. Tishby, Phys. A (Amsterdam, Neth.) **302**, 89 (2001).

[35] R. G. James, C. J. Ellison, and J. P. Crutchfield, Chaos **21**, 037109 (2011).

[36] S. E. Marzen and J. P. Crutchfield, Phys. Lett. A **380**, 1517 (2016).

[37] S. Marzen and J. P. Crutchfield, J. Stat. Phys. **168**, 109 (2017).

[38] S. E. Marzen and J. P. Crutchfield, J. Stat. Phys. **169**, 303 (2017).

[39] S. E. Marzen, M. R. DeWeese, and J. P. Crutchfield, Front. Comput. Neurosci. **9**, 105 (2015).

[40] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Probability and its Applications (Springer-Verlag, New York, 2003).

[41] R. E. Spinney, J. T. Lizier, and M. Prokopenko, arXiv:1712.09715.

[42] R. Spinney and I. Ford, in *Nonequilibrium Statistical Physics of Small Systems*, edited by R. Klages, W. Just, and C. Jarzynski (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2013), pp. 3–56.