Entropy balance and information processing in bipartite and nonbipartite composite systems

Richard E. Spinney, Joseph T. Lizier, and Mikhail Prokopenko

Complex Systems Research Group, School of Civil Engineering, Faculty of Engineering and IT,
The University of Sydney, Sydney, New South Wales, Australia, 2006

Information dynamics is an emerging description of information processing in complex systems that describes systems in terms of intrinsic computation, identifying computational primitives of information storage and transfer. In this paper we make a formal analogy between information dynamics and stochastic thermodynamics that describes the thermal behavior of small irreversible systems. As stochastic dynamics is increasingly being utilized to quantify the thermodynamics associated with the processing of information we suggest such an analogy is instructive, highlighting that existing thermodynamic quantities can be described solely in terms of extant information theoretic measures related to information processing. In this contribution we construct irreversibility measures in terms of these quantities and relate them to the physical entropy productions that characterize the behavior of single and composite systems in stochastic thermodynamics illustrating them with simple examples. Moreover, we can apply such a formalism to systems that do not have a bipartite structure. In particular we demonstrate that, given suitable nonbipartite processes, the heat flow in a subsystem can still be identified, and one requires the present formalism to recover generalizations of the second law. In these systems residual irreversibility is associated with neither subsystem and this must be included in the generalized second laws. This opens up the possibility of describing all physical systems in terms of computation allowing us to propose a framework for discussing the reversibility of systems traditionally out of scope of stochastic thermodynamics.

I. INTRODUCTION

In the early 1990s Evans and coworkers [1] first measured the probability of "second law violations" that occur in small systems. This newfound ability to quantify thermodynamic quantities at scales where uncertainties dominate opened up a new frontier in modern thermodynamics leading, notably, to the now famous work relations of Crooks and Jarzynski [2–5]. Later, a unifying framework for general stochastic systems [6] has formed the basis for a now indispensable tool for the study of small noisy systems, named *stochastic thermodynamics* [7,8]. More recently, however, such formalisms have implicated information theoretic measures introduced in contexts such as the exploitation of pure order [9–13], or more generally where feedback and/or measurement, is occurring [14–25] providing powerful relationships between physical quantities and information. A concerted effort is now being levelled at utilizing such a connection to understand the thermodynamics of computation [26–31].

Quite separately, however, there is an alternative viewpoint of physical (or otherwise) processes, which posits that they support *intrinsic computation* [32,33]. In this view, subsequent states of a dynamical process are considered to have been computed from earlier ones. In practice, however, the confidence in the precise values of such subsequent states is never perfect (arising through inherent coarse-graining, imprecise initial conditions, or in abstract, pure noise). Given this unavoidable reality, the central approach becomes the quantification of these uncertainties in a nonparametric manner through the introduction of information theory. This in turn provides a natural viewpoint for the characterization of the information processing supported by the dynamical process. In this view, the exact structure of physical dynamics gives way to the phenomenological predictive capacities and probabilities derived from them, from which, in turn, the structure of the information processing performed by the system can be characterized.

A. Background

Framing such a discussion is the concept of generalized computation and a challenge put forward by Langton [34,35] in the early 1990s: How can emergence of computation and, in particular, the emergence of complex spatiotemporal dynamics capable of universal computation, be explained in a dynamic setting? In doing so, he suggested to dissect computation into three primitive functions: the *transmission*, *storage*, and *modification of information*. Over the past decade, this decomposition has been formalized through an emerging field called information dynamics [36–46], proving itself to be useful in revealing or explaining generic mechanisms underlying dynamics of distributed information processing.

A common feature in several studies that used the primitive information-processing functions in tracing information dynamics is the distributed but coherent nature of the underlying computation [47]. For example, the coherent behavior given by swarms of animals that self-organize in complex large-scale spatial patterns is often explained by "collective memory" [48] and "collective communications" expressed via wave-like cascades of information transfer [49]. Interestingly, the information cascades have been conjectured to correspond to long range communications that either dynamically reorganize the swarm reducing the "fragility of mass behavior" [50] or propagate incorrect decisions [51].

It has been conjectured that information plays a dynamic role in such a self-organization [52], and more specifically, that distributed information-processing enables the groups to collectively perform computation [53–55]. These conjectures have been formalized in the context of information dynamics [56], verifying the hypothesis that the collective memory within a swarm can be captured by *active information storage*, while the information cascades are captured by *conditional transfer entropy* [39,41]. This has been further verified for real biological groups, such as swarms of soldier crabs [57], schools of zebrafish responding to a robotic animal replica [58], pairs of bats [59], rummy-nose tetras (*Hemigrammus rhodostomus*) fish schools [60], and so on.

A more abstract set of examples is given by dynamics of random Boolean networks—a canonical model for Gene Regulatory Networks (GRNs) [61]. When the average connectivity or activity level is varied, the dynamics undergo a phase transition between order and chaos, and it has been shown that the information storage and information transfer are maximized on either side of the critical point: For fully random topologies the information storage dominates the ordered phase of dynamics and the information transfer is the primary information-processing primitive at the chaotic phase [62]. When the underlying topology changes as well, specifically when it undergoes an order—small-world—randomness transition, the interplay between information storage and transfer attains a balance so that the network near the small-world state retains a sufficient amount of storage capability while being able to transfer information at a high rate [63].

Information dynamics methods have also been successfully used for online machine learning in robotic and artificial life scenarios [64–68], again allowing the designers to guide the automated learning process along specific information-processing primitives and fine-tune the computational balances, in addition to providing novel insights across a broad range of fields, including in canonical complex systems such as cellular automata [36], interpretation of dynamics in [43] and improved algorithms for machine learning [46], characterizing information processing signatures in biological signaling networks [44], nonlinear time-series forecasting [69], and in computational neuroscience applications in identifying neural information flows from brain imaging data [45,70,71], inferring effective network structure [72–74], providing evidence for the predictive coding hypothesis [75], and identifying differences in predictive information in autism spectrum disorder subjects and controls [76].

Despite these developments, the character of physical entropic balances between the information-processing primitives engaged in a distributed computation has not been fully elucidated. And indeed, until now the separate perspectives of stochastic thermodynamics and information dynamics have not been comprehensively understood in terms of each other.

As such our study has the following motivation—How might the computational primitives described in information dynamics manifest themselves in thermodynamics?—and in what way might physical bounds or constraints upon the system be expressed solely in terms of the computational primitives that the system supports? And do such quantities point to a dynamic generalization of the current thermodynamic bounds

purely in terms of information theoretic characterizations of the dynamics?

B. Outline

In this paper, to answer the above questions, we explicitly introduce the notion of irreversibility into the language of distributed computation and in doing so make a formal connection between information dynamics and stochastic thermodynamics, specifically in terms of stored and transferred information and the bounds they place on each other.

Through the definition of an appropriate *time-reversed computation*, we discuss the storage and transfer of information of time-reversed behavior and consequently identify irreversibility measures associated with stored and transferred information. Importantly, for physical systems that permit identification of an unambiguous entropy production, this total entropy production is identically equal to the sum of irreversibilities associated with storage and transfer of information. This allows us to identify contributions to the physical entropy production solely in terms of computational primitives.

We illustrate the behavior of such entropy production contributions in a variety of increasingly broad situations. First, we consider the canonical situation in stochastic thermodynamics, namely, that of a system controlled by an external protocol, illustrating that the entropy production associated with information transfer to the protocol is intimately related to the irreversibility in the dynamics of that protocol, demonstrating different bounds on the contributions for reversible and irreversible protocol dynamics. We then apply the framework to composite systems, considering first bipartite systems. We show that our framework identifies the physical entropy contributions that allow the second law to be locally broken for individual subsystems, for instance, in the context of feedback, in contrast to previous approaches and results which aim to bound such terms with information theoretic measures. In considering such composite systems, we introduce a principle of conserved predictive capacity, which constrains the extent to which computational primitives of the total system can be associated with the individual subsystems. We find that such a principle always leads to the existence of an interaction term understood to be a predictive capacity not uniquely attributable to any individual subsystem. Further, this interaction term is essential for constructing generalizations of the second law in terms of stored and transferred information. For bipartite systems this interaction term reduces to the change in mutual information recovering previous results where such terms are introduced *ad hoc* to construct the appropriate bounds.

Moreover, because our formalism identifies contributions to the entropy production using physically agnostic information theoretic measures, the resultant relations are valid where previous ones are not. Specifically, they apply to nonbipartite systems where we can still associate computational irreversibility despite unambiguous heat flows not necessarily being well defined. Importantly, in these systems the relations from our formalism that hold for bipartite systems remain *unchanged*, naturally accounting for intrinsic correlation in the dynamics of the subsystems. We emphasize that systems that are physically meaningful are not restricted to the bipartite case, yet previously no theory of how entropy production should be

associated with the behavior of individual subsystems has been provided. As such we posit that such a framework is a natural approach to the phenomena of irreversibility, applicable to all distributed computing systems, from which known bounds in stochastic thermodynamics emerge where the relevant physical assumptions hold.

We illustrate our results on two simple models. The first is a simple bipartite model allowing illustration of both entropy associated with storage and transfer of information. The second is a nonbipartite linear model of a feedback controller with system and controller driven by correlated noise. Until now we are aware of no formalism describing the information thermodynamics of this setup. We emphasize that physical heats are identifiable in such a system, that previous generalizations of the second law with information theoretic terms fail in this context, but the bounds from the present formalism do hold for any degree of correlation in the dynamics.

II. INTRINSIC COMPUTATION AND INFORMATION DYNAMICS

The outlook that intrinsic computation may be decomposed into distinct primitives naturally frames the discussion of computation in terms of identifying precise measures of information processing which information dynamics [36–46] aims to quantify.

Originally formulated in the context of distributed computing systems, possessing a characteristic update interval, in discrete time, the central object in information dynamics is the *predictive capacity* which quantifies the transformation of uncertainty in the "output," some next state in a given time series, when all previous "input" states, all current and previous time series data, are known. This predictive capacity, measured in bits or nats, is characterized by a mutual information, defined between two variables A and B taking values $a \in \mathcal{A}$ and $b \in \mathcal{B}$, respectively, as

$$
\begin{aligned}
I_M(A; B) &= \sum_{\substack{a \in \mathcal{A} \\ b \in \mathcal{B}}} p(A = a, B = b) \ln \frac{p(A = a, B = b)}{p(A = a)p(B = b)} \\
&= \left\langle \ln \frac{p(A = a, B = b)}{p(A = a)p(B = b)} \right\rangle,
\end{aligned} \tag{1}
$$

such that $\langle \ldots \rangle$ indicates an ensemble expectation. Specifically, the predictive capacity is the mutual information between the complete past of the universe and the next computed state of the variable in question. We characterize the evolution of the variable in question, X, taking values $x \in \mathcal{X}$, as a time series $X_{0:i} \equiv \{X_0, X_1, \ldots, X_i\}$, taking values $x_{0:i} \equiv \{x_0, x_1, \ldots, x_i\}$ in a space $\mathcal{X}_{0:i} = \mathcal{X}^{\otimes(i+1)}$, such that $x_i = x(t = i)$. Similarly, the rest of the universe is captured by an extraneous variable Y taking values in $y \in \mathcal{Y}$ which equally is characterized as a time series $Y_{0:i} \equiv \{Y_0, Y_1, \ldots, Y_i\}$, taking values $y_{0:i} \equiv \{y_0, y_1, \ldots, y_i\}$ in a space $\mathcal{Y}_{0:i} = \mathcal{Y}^{\otimes(i+1)}$. As such we consider the predictive capacity, $C_X^{\{i,i+1\}}$, from time $t = i$ to $t = i + 1$, for a process started at time $t = 0$ as

$$
\begin{aligned}
C_X^{\{i,i+1\}} &= I_M(X_{i+1}; X_{0:i} Y_{0:i}) \\
&= H(X_{i+1}) - H(X_{i+1}|X_{0:i}, Y_{0:i}).
\end{aligned} \tag{2}
$$

The superscripts indicate, explicitly, that this prediction occurs over the interval $\{i, i + 1\}$. In this sense we understand that it characterizes the total reduction in uncertainty that can be achieved about the "computed state," $X_{i+1} = x_{i+1}$, given the inputs, namely the previous values of the time series, $\{X_{0:i}, Y_{0:i}\} = \{x_{0:i}, y_{0:i}\}$. The information theoretic quantities are then built from ensemble probabilities $p(X_i = x_i)$ and transition probabilities $p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i})$ such that

$$
H(X_{i+1}) = - \sum_{x_{i+1} \in \mathcal{X}_{i+1}} p(X_{i+1} = x_{i+1}) \ln p(X_{i+1} = x_{i+1})
$$

$$
\begin{aligned}
&H(X_{i+1}|X_{0:i}, Y_{0:i}) \\
&= - \sum_{\substack{x_{0:i+1} \in \mathcal{X}^{\otimes(i+2)} \\ y_{0:i} \in \mathcal{Y}^{\otimes(i+1)}}} p(X_{0:i+1} = x_{0:i+1}, Y_{0:i} = y_{0:i}) \\
&\quad \times \ln p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i}, Y_{0:i} = y_{0:i}), \tag{3}
\end{aligned}
$$

etc., where here and throughout we will assume time homogeneity such that any and all time variation in the dynamics/transition probabilities experienced by X is parametrized through Y.

The insight from information dynamics is to divide this total predictive capacity into computationally relevant quantities in the spirit of Langton's conception of generalized computation. The pertinent division is that which separates the predictive capacity due to the process X and the predictive capacity over and above that of X due to Y. This is written

$$
\begin{aligned}
C_X^{\{i,i+1\}} &= H(X_{i+1}) - H(X_{i+1}|X_{0:i}) \\
&\quad + H(X_{i+1}|X_{0:i}) - H(x_{i+1}|X_{0:i}, Y_{0:i}) \\
&= I_M(X_{i+1}; X_{0:i}) + I_M(X_{i+1}; Y_{0:i}|X_{0:i}) \\
&= A_X^{\{i,i+1\}} + T_{Y \to X}^{\{i,i+1\}}, \tag{4}
\end{aligned}
$$

where $I_M(A; B|C) = I_M(A; B, C) - I_M(A; C)$ is a conditional mutual information between A and B given C. Here $A_X^{\{i,i+1\}}$ is known as the *active information storage* [42] and $T_{Y \to X}^{\{i,i+1\}}$ as the *transfer entropy* [77], a well known nonparametric measure of predictability used in many areas of science, in this instance from Y to X. We denote this division of the predictive capacity, the *computational signature* of the process (such a division has been explored previously; see, for example, Ref. [47]). The active information storage is interpreted as the stored information in X which is utilized in making a transition, while the transfer entropy is interpreted as the transferred information from Y utilized in making a transition in the context of the full history of X. Importantly, since both A_X and $T_{Y \to X}$ can be written as (conditional) mutual informations they are thus rigorously nonnegative.

Recent developments have stressed that these quantities should be understood as expectation values of suitable specific, pointwise or *local*, quantities [42,78]. These are known as the *local active information storage* a_x and *local transfer entropy* $t_{y \to x}$, where we write

$$
\begin{aligned}
A_X^{\{i,i+1\}} &= \langle a_x^{\{i,i+1\}} \rangle, \\
T_{Y \to X}^{\{i,i+1\}} &= \langle t_{y \to x}^{\{i,i+1\}} \rangle, \tag{5}
\end{aligned}
$$

with

$$a_x^{\{i,i+1\}} = \ln \frac{p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i})}{p(X_{i+1} = x_{i+1})},$$

$$t_{y \to x}^{\{i,i+1\}} = \ln \frac{p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i}, Y_{0:i} = y_{0:i})}{p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i})}, \quad (6)$$

which together comprise the local predictive capacity

$$c_x^{\{i,i+1\}} = a_x^{\{i,i+1\}} + t_{y \to x}^{\{i,i+1\}}$$

$$= \ln \frac{p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i}, Y_{0:i} = y_{0:i})}{p(X_{i+1} = x_{i+1})}, \quad (7)$$

with $C_X^{\{i,i+1\}} = \langle c_x^{\{i,i+1\}} \rangle$. In purely information theoretic terms, such local quantities may be interpreted as the differences in code lengths between competing models of the observed behavior in X_{i+1}. 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. Detailed understanding of these *individual* realizations using such local quantities has yielded important identification and insights into distributed computing behavior. This ability to attribute storage and transfer of information to individual realizations is crucial to the following developments.

III. TIME-REVERSED COMPUTATIONS

Modern thermodynamics makes explicit links between the dynamic irreversibility of a physical process and the total entropy production of the universe. Indeed, this "total entropy production," comprising the change in internal uncertainty and the heat dissipated to an environment, can be, under the right circumstances, considered equal to the log-ratio of the probability of observing the "forward" behavior relative to the probability of observing the "reverse" behavior, having started from an ensemble characterized by the distribution at the end of the forward process. This entails, if not the construction, the consideration of a suitable "reverse process" with which the normal, forward dynamics are contrasted.

We wish to give an alternative account of irreversibility in *computational* terms, understood in the sense of information dynamics. To do so we need to attribute computational quantities to the time-reverse behavior of the system. Accordingly, and in contrast to the reverse process considered in stochastic thermodynamics, we define a *time-reversed computation* through an analogous *time-reversed predictive capacity*. Such an object is defined axiomatically, on a local, pointwise, scale, as the predictive capacity with all constitute elements time-reversed relative to the transition being considered in the context of the whole process from t_0 to t_n, viz.

$$c_x^{\dagger,\{n-(i+1),n-i\}} = \ln \frac{p^\dagger(X_{i+1} = x_{i+1}^\dagger|X_{0:i} = x_{0:i}^\dagger, Y_{0:i} = y_{0:i}^\dagger)}{p^\dagger(X_{i+1} = x_{i+1}^\dagger)}, \quad (8)$$

noting that this has been defined for the transition from time $n-i-1 \to n-i$. Here we have introduced

(1) The time-reversed transform denoted by the $\dagger$ symbol which time reverses and reflects sequences on the interval $[0, n]$. This leads to the time-reversed path $x_{0:n}^\dagger$, which in the absence of odd parity variables amounts to the reverse sequence of $x_{0:n}$, i.e., $x_{0:n}^\dagger = x_{n:0} = \{x_n, \ldots, x_0\}$, such that $x_i^\dagger = x_{n-i}$.

(2) The time-reversed ensemble probability, appearing in the denominator, defined as $p^\dagger(X_i = x_i^\dagger) = p(X_i^\dagger = x_i^\dagger) = p(X_{n-i} = x_{n-i})$.

(3) The *time-reversed dynamics* [6,79,80], appearing in the numerator, which in contrast accounts for any differences that occur under time reversal in the generator of the process. This can arise due to dependence on odd quantities with respect to time reversal (e.g., magnetic fields/torques etc.) or differences in evolution related to intervention in the system, e.g., in extraneous variables Y. Unless otherwise stated we will consider $p^\dagger = p$ relating to autonomous, time invariant, behavior.

We emphasize that the probability in the numerator is of the reverse trajectory as if it were evolving forward in time under dynamics denoted $p^\dagger$. In this sense it remains a *prediction* of a hypothetical future evolution of the system. We emphasize that $p^\dagger$ is *not* a probability associated with *retrodiction*, i.e., *not* $p^\dagger(X_{i+1}^\dagger = x_{i+1}^\dagger|X_{0:i}^\dagger = x_{0:i}^\dagger, Y_{0:i}^\dagger = y_{0:i}^\dagger)$.

For brevity we subsequently drop the explicit reference to the variables and write the probabilities solely in terms of the realized values such that we write

$$c_x^{\{i,i+1\}} = \ln \frac{p(x_{i+1}|x_{0:i}, y_{0:i})}{p(x_{i+1})},$$

$$c_x^{\dagger,\{n-i-1,n-i\}} = \ln \frac{p^\dagger(x_{i+1}^\dagger|x_{0:i}^\dagger, y_{0:i}^\dagger)}{p^\dagger(x_{i+1}^\dagger)}$$

$$= \ln \frac{p^\dagger(x_{n-i-1}|x_{n:n-i}, y_{n:n-i})}{p(x_{n-i-1})}, \quad (9)$$

and so, for the transition $i \to i+1$,

$$c_x^{\dagger,\{i,i+1\}} = \ln \frac{p^\dagger(x_i|x_{n:i+1}, y_{n:i+1})}{p(x_i)}, \quad (10)$$

where we may understand $p(x_i) \equiv p(X_i = x_i)$, $p(x_{i+1}|x_i) \equiv p(X_{i+1} = x_{i+1}|X_i = x_i)$, $p^\dagger(x_i|x_{i+1}) \equiv p^\dagger(X_{n-i} = x_i|X_{n-i-1} = x_{i+1})$, and so on. We also note that by time homogeneity we may also write $p^\dagger(X_{n-i} = x_i|X_{n-i-1} = x_{i+1}) = p^\dagger(X_{i+1} = x_i|X_i = x_{i+1})$. Some of these quantities, in the context of continuous time, are illustrated in Fig. 1. Here we observe two path realizations $x_{[t_0,\tau]} \equiv \{x(t), t \in [t_0,\tau]\}$ and $y_{[t_0,\tau]} \equiv \{y(t), t \in [t_0,\tau]\}$ of X and Y, defined between the time origin t_0 ($t = 0$ in discrete time) and time horizon τ ($t = n$ in discrete time) illustrated in full in the first subplot. The second subplot concerns the prediction associated with the intrinsic computation at a given time t, which can be understood as the midpoint between times i and $i+1$ in discrete time. When X is the computing quantity the predictive capacity quantifies the ability to predict the next step along the dashed blue line given knowledge of the solid blue and red lines. The capacity derived from the dark line is the active information storage whilst the capacity derived from the light line is the transfer entropy. The third subplot concerns the time-reversed prediction associated with the
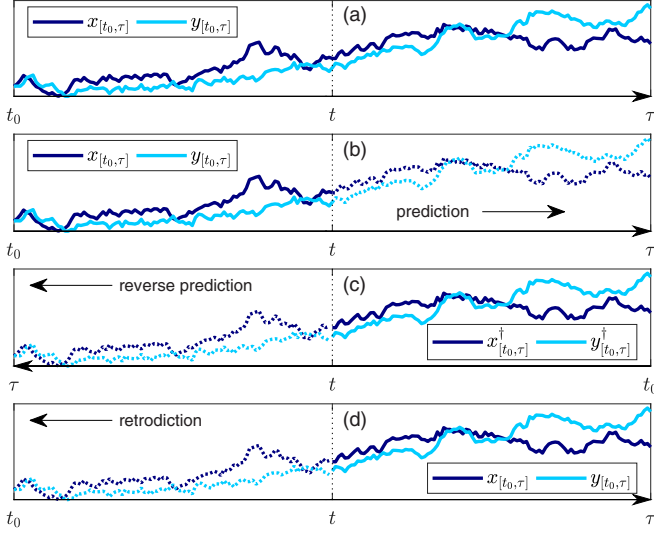
FIG. 1. Illustration of the intrinsic and reverse computations and predictions. In subfigure (a) we plot two paths associated with random functions $X_{[t_0,\tau]}$ and $Y_{[t_0,\tau]}$ with time origin and horizon t_0 and τ, with a specific intermediate point, $t_0 < t < \tau$, being considered. At this time the forward (ensemble) computation in X concerns the prediction of the dashed dark line on the right-hand side in subfigure (b) from the solid dark and light lines on the left-hand side using the ensemble dynamics p. The reverse computation in X concerns the prediction of the dashed dark line in subfigure (c) from right to left using the time-reversed dynamics $p^{\dagger}$. This is to be contrasted with retrodiction illustrated in subfigure (d).

time-reversed computation. Here the reverse sequence of X is being predicted about time t; however, the prediction is in the same direction as the time evolution as the physical, time-reversed dynamics are being utilized. The fourth subplot serves to distinguish the reverse prediction from a *retrodiction*: in this plot, *previous* values of X are being retrodicted after the fact from the subsequent trajectories.

Importantly, as with the intrinsic computation, the time-reversed predictive capacity naturally divides into two components:

$$c_x^{\dagger,\{i,i+1\}} = \ln \frac{p^{\dagger}(x_i | x_{n:i+1}, y_{n:i+1})}{p(x_i)}$$
$$= \ln \frac{p^{\dagger}(x_i | x_{n:i+1})}{p(x_i)} + \ln \frac{p^{\dagger}(x_i | x_{n:i+1}, y_{n:i+1})}{p^{\dagger}(x_i | x_{n:i+1})}. \quad (11)$$

The first characterizes the uncertainty reduction due to the history of the variable which defines the computed quantity, X, and the second characterizes uncertainty reduction due to the extraneous time series, Y. The second quantity exists in the literature and is the local *time-reversed transfer entropy* [81],

$$t_{y \to x}^{\dagger,\{i,i+1\}} = \ln \frac{p^{\dagger}(x_i | x_{n:i+1}, y_{n:i+1})}{p^{\dagger}(x_i | x_{n:i+1})}, \quad (12)$$

which becomes a pathwise quantity when integrated along a path [82]. The first quantity, however, has not previously been addressed and is a quantity we introduce which we call the

local *time-reversed active information storage*

$$a_x^{\dagger,\{i,i+1\}} = \ln \frac{p^{\dagger}(x_i | x_{n:i+1})}{p(x_i)}. \quad (13)$$

Together, these quantities characterize the distributed structure of time-reversed computation and crucially do so on a *local* scale for individual realizations of the process.

IV. A MEASURE OF COMPUTATIONAL IRREVERSIBILITY

In quantifying the irreversibility of the computation we seek to contrast the computational signature of the intrinsic computation with the computational signature of the time-reversed computation. Further, any such measure should characterize individual realizations of a process. We identify such a quantity as the difference between computational signatures, characterizing, on the local level, the difference in predictive capacity between the two predictions/computations. We introduce such a quantity, which we call the *computational irreversibility*, as follows:

$$\mathcal{I}_x^{\{i,i+1\}} = c_x^{\{i,i+1\}} - c_x^{\dagger,\{i,i+1\}}$$
$$= a_x^{\{i,i+1\}} - a_x^{\dagger,\{i,i+1\}} + t_{y \to x}^{\{i,i+1\}} - t_{y \to x}^{\dagger,\{i,i+1\}}, \quad (14)$$

where we identify components that contribute due to the predictive influence of Y and X as follows:

$$\mathcal{I}_x^{\{i,i+1\}} = \mathcal{I}_x^{A,\{i,i+1\}} + \mathcal{I}_{x \leftarrow y}^{T,\{i,i+1\}}$$
$$\mathcal{I}_x^{A,\{i,i+1\}} = a_x^{\{i,i+1\}} - a_x^{\dagger,\{i,i+1\}} \quad (15)$$
$$\mathcal{I}_{x \leftarrow y}^{T,\{i,i+1\}} = t_{y \to x}^{\{i,i+1\}} - t_{y \to x}^{\dagger,\{i,i+1\}}.$$

As such we identify $\mathcal{I}_x^{A,\{i,i+1\}}$ as *intrinsic* computational irreversibility associated with *storage* of information in the transition from time i to $i+1$ and $\mathcal{I}_{x \leftarrow y}^{T,\{i,i+1\}}$ as *extrinsic* computational irreversibility associated with the *transfer* of information over that transition. Just as the active information storage quantifies the total predictability that can be extracted from the history of a time series, the intrinsic irreversibility quantifies the total irreversibility that can be identified using only that time series. Analogously, as the transfer entropy quantifies the additional predictability an extraneous time series provides, the extrinsic irreversibility quantifies the additional irreversibility that arises when Y is known in addition to X.

To proceed we need to associate such quantities with complete paths. As formulated, information dynamics considers its measures of storage and transfer to be associated with individual time steps. Consequently, we consider the difference in *cumulatively* utilized information storage and information transfer along entire paths by considering the sum of those individual storage and transfer measures associated with each time step, here from a time origin $t = 0$ to a time horizon $t = n$ obtaining, first, for the intrinsic irreversibility

$$\Delta \mathcal{I}_x^{A,\{0,n\}} = \sum_{i=0}^{n-1} \mathcal{I}_x^{A,\{i,i+1\}}$$
$$= \sum_{i=0}^{n-1} a_x^{\{i,i+1\}} - a_x^{\dagger,\{i,i+1\}}$$

$$= \sum_{i=0}^{n-1} \ln \frac{p(x_{i+1}|x_{0:i})}{p(x_{i+1})} \frac{p(x_i)}{p^{\dagger}(x_i|x_{n:i+1})}$$

$$= \ln \frac{p(x_0)}{p(x_n)} + \sum_{i=0}^{n-1} \ln \frac{p(x_{i+1}|x_{0:i})}{p^{\dagger}(x_i|x_{n:i+1})}$$

$$= \ln \frac{p(x_{0:n})}{p^{\dagger}(x_{n:0})}. \qquad (16)$$

Similarly the cumulative extrinsic irreversibility is defined as

$$\Delta \mathcal{I}_{x \leftarrow y}^{T,\{0,n\}} = \sum_{i=0}^{n-1} \mathcal{I}_{x \leftarrow y}^{T,\{i,i+1\}}$$

$$= \sum_{i=0}^{n-1} t_{y \to x}^{\{i,i+1\}} - t_{y \to x}^{\dagger,\{i,i+1\}}$$

$$= \sum_{i=0}^{n-1} \ln \frac{p(x_{i+1}|x_{0:i}, y_{0:i})}{p(x_{i+1}|x_{0:i})} \frac{p^{\dagger}(x_i|x_{n:i+1})}{p^{\dagger}(x_i|x_{n:i+1}, y_{n:i+1})}$$

$$= \ln \frac{p^{\dagger}(x_{n-1:0}|x_n)}{p(x_{1:n}|x_0)} \frac{p(x_{1:n}|x_0, \{y_{0:n-1}\})}{p^{\dagger}(x_{n-1:0}|x_n, \{y_{n:1}\})}, \qquad (17)$$

where we utilize notation for probabilities of complete paths as follows:

$$p(x_{1:n}|x_0) = \prod_{i=1}^{n} p(x_i|x_{0:i-1}),$$

$$p^{\dagger}(x_{n-1:0}|x_n) = \prod_{i=0}^{n-1} p(x_i|x_{n:i+1}),$$

$$p(x_{0:n}) = p(x_0)p(x_{1:n}|x_0),$$

$$p^{\dagger}(x_{n:0}) = p(x_n)p(x_{n-1:0}|x_n),$$

$$p(x_{1:n}|x_0, \{y_{0:n-1}\}) = \prod_{i=0}^{n-1} p(x_{i+1}|x_{0:i}, y_{0:i}),$$

$$p^{\dagger}(x_{n-1:0}|x_n, \{y_{n:1}\}) = \prod_{i=0}^{n-1} p(x_i|x_{n:i+1}, y_{n:i+1}),$$

(18)

with notation {} emphasizing that $p(x_{1:n}|x_0, \{y_{0:n-1}\}) \neq p(x_{1:n}|x_0, y_{0:n-1})$ as future values of Y are causally blocked along evaluation of the path.

Explicitly, $\Delta \mathcal{I}_x^{A,\{0,n\}}$ is the difference between the cumulative use of stored information utilized in a specific path for all computational steps from $t = 0$ through to $t = n$ and the cumulative use of stored information in all steps of the appropriate reverse computation sequenced from n to 0. Analogously, $\Delta \mathcal{I}_{x \leftarrow y}^{T,\{0,n\}}$ is the difference between the cumulatively transferred information in the forward computations and the cumulatively transferred information in the appropriate reverse computations. Both of these are expressible as probabilities assigned to total path realizations, i.e., $x_{0:n}$, etc.

Importantly, these accumulated quantities along such paths can be expressed as functionals consisting of ratios of the constructed path probabilities. This supports their existence when a

limit in the discretization is taken in the approach to continuous time (thus permitting a rate). In the continuum regime we are concerned with such quantities formulated for continuous path functions written $x_{\mathcal{A}} \equiv \{x(t') : t' \in \mathcal{A}\}$. We construct these quantities by considering an interval $[t_0, \tau]$ and constructing a discretization running from $i = 0$ to $i = n = (\tau - t_0)/\Delta t$ such that $x_i = x(t_0 + i\Delta t)$ then taking $\Delta t \to 0$ such that $x_{0:n} \to x_{[t_0,\tau]}$, $x_{1:n} \to x_{(t_0,\tau]}$, and so on. In this limit, measures replace probabilities and computational signatures and irreversibilities become differential quantities defined instantaneously; i.e., $\mathcal{I}_x^{\{i,i+1\}} \to d\mathcal{I}_x(t)$. We consequently identify the accumulated quantities with logarithms of Radon-Nikodym derivatives between these path measures [82,83], such that for a process running from $t = t_0$ to $t = \tau$ we write

$$\Delta \mathcal{I}_x^A = \int_{t_0}^{\tau} d\mathcal{I}_x^A(t) = \lim_{\Delta t \to 0} \ln \frac{p(x_{0:n})}{p^{\dagger}(x_{n:0})} = \ln \frac{dp[x_{[t_0,\tau]}]}{dp^{\dagger}[x_{[t_0,\tau]}]},$$

$$\Delta \mathcal{I}_{x \leftarrow y}^T = \int_{t_0}^{\tau} d\mathcal{I}_{x \leftarrow y}^T(t)$$

$$= \lim_{\Delta t \to 0} \ln \frac{p^{\dagger}(x_{n-1:0}|x_n)}{p(x_{1:n}|x_0)} \frac{p(x_{1:n}|x_0, \{y_{0:n-1}\})}{p^{\dagger}(x_{n-1:0}|x_n, \{y_{n:1}\})}$$

$$= \ln \frac{dp^{\dagger}[x_{(\tau,t_0]}|x_\tau]}{dp[x_{(t_0,\tau]}|x_{t_0}]} \frac{dp[x_{(t_0,\tau]}|x_{t_0}, \{y_{[t_0,\tau)}\}]}{dp^{\dagger}[x_{(\tau,t_0]}|x_\tau, \{y_{[\tau,t_0)}\}]}. \qquad (19)$$

For clarity, however, we utilize a deliberate abuse of notation rewriting measures as path probability "densities" such that

$$\int dp[x_{[t_0,\tau]}] \equiv \int dx_{[t_0,\tau]} \, p[x_{[t_0,\tau]}]$$

$$= \int dx_{[t_0,\tau]} \, p[x_{(t_0,\tau]}|x_{t_0}]p(x_{t_0}), \qquad (20)$$

since this permits a more natural discussion and intuitive manipulation when dealing with familiar objects such as mutual informations, etc., without altering any of the results.

Before progressing we note that the intrinsic irreversibility associated with information storage permits a nonnegative ensemble average viz.

$$\langle \Delta \mathcal{I}_x^A \rangle = \int dx_{[t_0,\tau]} \, p[x_{[t_0,\tau]}] \ln \frac{p[x_{[t_0,\tau]}]}{p^{\dagger}[x_{[\tau,t_0]}]} \geqslant 0, \qquad (21)$$

but no bound exists for $\langle \Delta \mathcal{I}_x \rangle$ or $\langle \Delta \mathcal{I}_{x \leftarrow y}^T \rangle$, properties that we shall return to.

V. COMPUTATIONAL IRREVERSIBILITY IN PHYSICAL SYSTEMS

The measures $\langle \Delta \mathcal{I}_x \rangle$, $\langle \Delta \mathcal{I}_x^A \rangle$, and $\langle \Delta \mathcal{I}_{x \leftarrow y}^T \rangle$ are very general objects which quantify irreversibility for any dynamics where the relevant probability measures can be defined. However, to understand their potential role in, and connection with, (stochastic) thermodynamics we must consider their behavior in relevant physical systems. This will necessarily involve a constraining of the nature of the systems considered. To understand the relevant constraints it is instructive to give a

brief account of how irreversibility in physical systems can be identified thermodynamically as entropy production.

A. Entropy production in stochastic thermodynamics

Here we give a brief account of how entropy production can be understood in terms of dynamic irreversibility in systems that obey such Markovian dynamics, thus giving a brief overview of the necessary features of stochastic thermodynamics. As with all thermodynamic descriptions, we consider a division of the universe into a system of interest with degrees of freedom that we are aware of and an environment that we are mostly unconcerned with and so remain ignorant of. If the system is small enough, this ignorance of the environment renders the behavior of the system unpredictable, or stochastic due to our failure to track its precise evolution. Further, if we have performed such a division sensibly, the environmental degrees of freedom will be fast in comparison to the system allowing us to make an assumption that the induced probabilistic dynamics are Markov, but also the assumption that the environment is at equilibrium. In such situations a powerful principle arises known as *local detailed balance* [84–91], which relates the transition rates that arise from exposure to the environment to the exported entropy production realized in the environment.

If we consider such a continuous time Markovian process in X on a discrete state space we may describe its behavior as a sequence of instantaneous transitions. On an interval $[t_0, \tau]$ there is a countable number of transitions, N, occurring at times $t_0 < t_1 < \ldots < t_N < \tau$ between states $X_t = x_i$, $t \in [t_i, t_{i+1})$. The dynamics are consequently described by a set of transition rates $k^y_{x \to x'}$ concerning a transition from x to x', which we also allow to be parametrized by a smoothly varying switching protocol Y. In such a scenario any such transition is mediated by the equilibrium environment and thus the local detailed balance relation stipulates that it is accompanied by an entropy change in the environment, at time t_i, equal to (setting $k_B = 1$ here and throughout)

$$\ln \frac{k^{y_{t_i}}_{x_{i-1} \to x_i}}{k^{y_{t_i}}_{x_i \to x_{i-1}}} = \Delta \mathcal{S}^x_{\text{env}}(t_i), \tag{22}$$

manifesting as a heat flow scaled by the environmental temperature when the environment is a heat bath. By definition, the probability of observing such a transition in a given time Δt is then written $p(X_{t+\Delta t} = x' | X_t = x, Y_t = y) = k^y_{x \to x'} \Delta t + O(\Delta t^2)$ such that we can associate the above with transition probabilities on a short timescale,

$$\Delta \mathcal{S}^x_{\text{env}}(t_i)$$
$$= \ln \frac{k_{x_{i-1} \to x_i}}{k_{x_i \to x_{i-1}}}$$
$$= \lim_{\Delta t \to 0} \ln \frac{p(X_{t_i} = x_i | X_{t_i - \Delta t} = x_{i-1}, Y_{t_i - \Delta t} = y_{t_i - \Delta t})}{p(X_{t_i} = x_{i-1} | X_{t_i - \Delta t} = x_i, Y_{t_i - \Delta t} = y_{t_i})}. \tag{23}$$

If we have invariance in the time-reversed dynamics $p^\dagger = p$ (i.e. $k^y_{x \to x'} = k^{y, \dagger}_{x \to x'}$) and Y is suitably smooth then we may

rewrite the above as

$$\Delta \mathcal{S}^x_{\text{env}}(t_i)$$
$$= \lim_{\Delta t \to 0} \ln \frac{p(X_{t_i} = x_i | X_{t_i - \Delta t} = x_{i-1}, Y_{t_i - \Delta t} = y_{t_i - \Delta t})}{p^\dagger(X_{t_i} = x_{i-1} | X_{t_i - \Delta t} = x_i, Y_{t_i - \Delta t} = y_{t_i - \Delta t})}$$
$$= \lim_{\Delta t \to 0} \ln \frac{p(x_{t_i} | x_{t_i - \Delta t}, y_{t_i - \Delta t})}{p^\dagger(x_{t_i - \Delta t} | x_{t_i}, y_{t_i})}, \tag{24}$$

i.e., in a form represented by the time-reversed dynamics and path sequence.

In contrast, an absence of entropy production in the environment (due to X) is associated with the times between transitions, understood physically as an absence of exchange of energy with the environment. This too can be captured in terms of probabilities of the transition sequence by simply observing the probability of such an absence in the forward and time-reversed dynamics to be exponentially distributed due to the Markovian dynamics and thus given by

$$\ln \frac{p\left(X_{(t_i, t_{i+1})} = x_i | X_{t_i} = x_i, \{Y_{[t_i, t_{i+1})} = y_{[t_i, t_{i+1})}\}\right)}{p^\dagger\left(X_{(t_i, t_{i+1})} = x_i | X_{t_i} = x_i, \{Y_{[t_i, t_{i+1})} = y_{[t_{i+1}, t_i)}\}\right)}$$
$$= \sum_{x'} \int_{t_i}^{t_{i+1}} \left(k^{y(t), \dagger}_{x_i \to x'} - k^{y(t)}_{x_i \to x'}\right) dt, \tag{25}$$

which vanishes when $p = p^\dagger$ such that $k_{x \to x'} = k^\dagger_{x \to x'}$. These contributions can be constructed in a piecewise fashion, with waiting times interleaved between transitions, allowing us to consider the ratio of the conditional probability of the entire path $x_{[t_0, \tau]}$ and the conditional probability of the entire path $x_{[\tau, t_0]}$ under the time-reversed dynamics allowing us to express the total entropy exported by the system X to the environment, along an entire path $x_{[t_0, \tau]}$ as [6,92]

$$\Delta \mathcal{S}^x_{\text{env}}[x_{[t_0, \tau]}] = \int_{t_0}^\tau d\mathcal{S}^x_{\text{env}}(t) = \ln \frac{p[x_{(t_0, \tau]} | x_{t_0}, \{y_{[t_0, \tau)}\}]}{p^\dagger[x_{(\tau, t_0]} | x_\tau, \{y_{[\tau, t_0)}\}]}, \tag{26}$$

in this case with $d\mathcal{S}^x_{\text{env}}(t) = \delta(t - t_i) \Delta \mathcal{S}_{\text{env}}(t_i)$. While we have described the above for a jump process, the above relationship is very robust, holding for all Markov systems that possess a local detailed balance relationship. This can be understood by recognizing that continuous dynamics, realized, for instance, as stochastic differential equations, may be achieved through the appropriate limit of such a jump process (i.e., the Fokker Planck equation may be derived from the master equation).

Having established the form of the exported entropy production, one of the key insights of stochastic thermodynamics is to observe that when augmenting such a quantity with a change in *pointwise* (or *local*) Shannon entropy of the system, the sum then represents the total entropy production, in model, of the universe [6] (consisting of the system and the environment) and obeys the requisite statistical requirements to satisfy a second law. Such a change in Shannon entropy, for any class of system on the interval $[t_0, \tau]$, reads

$$\Delta \mathcal{S}^x_{\text{sys}}[x_{[t_0, \tau]}] = \ln \frac{p(x_{t_0})}{p(x_\tau)}. \tag{27}$$

If we consider such a change between the start and end of the process on $x_{[t, \tau]}$ this provides an initial condition for the path

probabilities appearing in Eq. (26) such that

$$
\begin{aligned}
&\Delta \mathcal{S}^x_{\mathrm{tot}}\big[x_{[t_0,\tau]}, y_{[t_0,\tau]}\big] \\
&= \Delta \mathcal{S}^x_{\mathrm{sys}}\big[x_{[t_0,\tau]}, y_{[t_0,\tau]}\big] + \Delta \mathcal{S}^x_{\mathrm{med}}\big[x_{[t_0,\tau]}, y_{[t_0,\tau]}\big] \\
&= \ln \frac{p\big[x_{[t_0,\tau]}|x_{t_0}, \{y_{[t_0,\tau)}\}\big] p(x_{t_0})}{p^\dagger\big[x_{[\tau,t_0]}|x_\tau, \{y_{[\tau,t_0)}\}\big] p(x_\tau)} = \ln \frac{p\big[x_{[t_0,\tau]}\big|\{y_{[t_0,\tau]}\}\big]}{p^\dagger\big[x_{[\tau,t_0)}\big|\{y_{[\tau,t_0)}\}\big]}.
\end{aligned}
$$
(28)

It is this quantity that the second law governs, and it does so on a statistical level. As such, assuming no feedback between the system X and switching protocol Y (i.e., Y cannot depend on X), the following holds:

$$
\langle \Delta \mathcal{S}^x_{\mathrm{tot}} \rangle = \langle \Delta \mathcal{S}^x_{\mathrm{sys}} \rangle + \langle \Delta \mathcal{S}^x_{\mathrm{env}} \rangle \geqslant 0,
$$
(29)

representing the second law.

Finally, however, it is important to note two key assumptions. The first is the assumption of no feedback. If this is not met then the second law as written above need not hold. To recover the second law in such cases either the definition of the system must be expanded or the result generalized. The second is the assumption of smoothness in Y in the preceding identification of the entropy production and its representation in terms of dynamical probabilities. If such a criterion is not met the notion of a heat physically dissipated by X can become fundamentally ambiguous since the value y that enters the local detailed balance condition is not well defined. This can occur notably if (i) both X and Y are continuous but nowhere differentiable (as in the case of Langevin equations) and driven by correlated noise or (ii) both X and Y are discrete variables which can simultaneously transition. In the latter, such a situation can be avoided by insisting that joint transitions are disallowed such that the dynamics of *both* X and Y, governed by the transition rates $k^{y \to y'}_{x \to x'}$ ($\{x, y\} \neq \{x', y'\}$), may be written

$$
k^{y \to y'}_{x \to x'} = (1 - \delta_{x,x'}) \delta_{y,y'} k^y_{x \to x'} + (1 - \delta_{y,y'}) \delta_{x,x'} k^{y \to y'}_x,
$$
(30)

such that $\sum_{y'} k^{y \to y'}_{x \to x'} = k^y_{x \to x'}$ and $\sum_{x'} k^{y \to y'}_{x \to x'} = k^{y \to y'}_x$. More broadly, such a requirement specifies conditional independence in the subsequent values of X and Y given the preceding history over small timescales, arising here since the transition times of X and Y are not correlated as they do not transition together, almost surely. This conditional independence is expressed by the following:

$$
\begin{aligned}
&\lim_{dt \to 0} p(x_{t+dt}, y_{t+dt}|x_t, y_t) \\
&= \lim_{dt \to 0} p(x_{t+dt}|x_t, y_t) p(y_{t+dt}|x_t, y_t),
\end{aligned}
$$
(31)

which we can observe to hold for the discrete space case by considering a transition $x \to x' \neq x$, $y = y'$, without loss of generality, such that we have

$$
\begin{aligned}
&p(X_{t+dt} = x'|X_t = x, Y_t = y) p(Y_{t+dt} = y|X_t = x, Y_t = y) \\
&= k^y_{x \to x'} dt \left(1 - \sum_{y'} k^{y \to y'}_x dt\right) + O(dt^2) \\
&= k^y_{x \to x'} dt + O(dt^2) \\
&= p(X_{t+dt} = x', Y_{t+dt} = y|X_t = x, Y_t = y) + O(dt^2).
\end{aligned}
$$
(32)

In the former case with continuous paths such a conditional independence requirement manifests as statistical independence of any noise terms between X and Y. When such a requirement is met then one may write

$$
\ln \frac{p(x_{t+dt}|x_t, y_t)}{p^\dagger(x_t|x_{t+dt}, y_{t+dt})} = \ln \frac{p(x_{t+dt}|x_t, y_t)}{p^\dagger(x_t|x_{t+dt}, y_t)} + O(dt^2),
$$
(33)

identifying a term representing the local detailed balance condition on the right-hand side and recovering Eq. (26).

This property of conditional independence, manifest as independence in noise terms in continuous dynamics and illegality of joint transitions in discrete dynamics, is called a *bipartite* property.

B. Relating computational irreversibility and physical entropy production

Here we make an explicit connection between the measures of computational irreversibility established in Sec. IV and the thermodynamic quantities described above. To do so we consider, first, a system where the physical assumptions listed above are true, namely, that (i) an equilibrium environment leads to stochastic dynamics in a system, X, which are Markov, thus leading to a local detailed balance relation, and that (ii) the joint dynamics of the system X and some extraneous parameter Y, which may represent an external protocol or some other coupled subsystem, are bipartite such that we observe no joint transitions or consider independence of noise in the continuous limit.

In this case we straightforwardly compare the contents of Eqs. (28) and (19) and immediately recognize a central observation:

$$
\Delta \mathcal{I}_x = \Delta \mathcal{S}^x_{\mathrm{tot}}.
$$
(34)

Explicitly, the computational irreversibility reduces to the total thermodynamic entropy production in the case of Markovian bipartite dynamics revealing that the central measure of intrinsic computation, the predictive capacity, can be seen to be an implicitly central quantity in stochastic thermodynamics. In other words, entropy production can be equivalently described by physical irreversibility and computational irreversibility as measured by information dynamics.

Since thermodynamic and computational irreversibility are equivalent in such a setting and the computational irreversibility is designed around a division in storage and transfer of information, it is instructive to translate such a division, realized in the intrinsic and extrinsic computation irreversibilities in Eq. (19), into divisions of physical entropy production. As such we therefore explicitly write

$$
d\mathcal{S}^x_{\mathrm{tot}} = d\mathcal{S}^x_{\mathrm{storage}} + d\mathcal{S}^{x \leftarrow y}_{\mathrm{transfer}},
$$
(35)

where we identify $d\mathcal{S}^x_{\mathrm{storage}} = d\mathcal{I}^A_x$ as the *physical entropy production* associated with the active storage of information and $d\mathcal{S}^{x \leftarrow y}_{\mathrm{transfer}} = d\mathcal{I}^T_{x \leftarrow y}$ as the physical entropy production associated with the transfer of information based on the identification $d\mathcal{S}^x_{\mathrm{tot}} = d\mathcal{I}_x = d\mathcal{I}^A_x + d\mathcal{I}^T_{x \leftarrow y}$. The last quantity, $d\mathcal{S}^{x \leftarrow y}_{\mathrm{transfer}}$, was used in the construction of a measure of information theoretic time asymmetry, an "information theoretic arrow of time" [81]. We emphasize here, however, that we are making a complete division of the entropy production associated with

the computational primitives, deriving from general measures of irreversibility.

Such a division is to be contrasted with other important divisions of the total entropy production of the universe, such as the usual division into the entropy change attributed to the system and the environment [6] (or in the language of Prigogine, the system into the internal and external productions [93]),

$$dS^x_{\text{tot}} = dS^x_{\text{sys}} + dS^x_{\text{env}}, \tag{36}$$

or the more recent division of the total entropy production into adiabatic and nonadiabatic contributions [92,94–97]:

$$dS^x_{\text{tot}} = dS^x_{\text{na}} + dS^x_{\text{a}}. \tag{37}$$

All of the above divisions of the total entropy production are based, in some sense, on how they manifest. The division in Eq. (36) characterizes the *location* associated with the entropy change: dS^x_{sys} concerns the entropy change associated with the system, while dS^x_{env} concerns the entropy change of the environment. The division in Eq. (37) characterizes the ensemble level *mechanisms* that lead to dissipation: dS^x_{na} concerns the entropy production associated with driving and relaxation and dS^x_{a} concerns the entropy production associated with steady nonequilibrium constraints. Our division in Eq. (35) does so based on the manifestation of computational behavior (storage and transfer), defined in terms of *predictive capacity* based on the terms laid out in Sec. II.

Finally, we mention that since the total computational irreversibility can be associated with physical entropy production, by assuming a heat bath at constant temperature β^{-1} (we use $k_B = 1$ throughout), we can write

$$dS^x_{\text{storage}} = dS_{\text{sys}} + d\beta Q^x_{\text{storage}},$$
$$dS^{x \leftarrow y}_{\text{transfer}} = d\beta Q^{x \leftarrow y}_{\text{transfer}}, \tag{38}$$

where

$$\Delta Q^x = \Delta Q^x_{\text{storage}} + \Delta Q^{x \leftarrow y}_{\text{transfer}}, \tag{39}$$

reminiscent of the identification of the nonadiabatic entropy production with the system entropy and excess heat [92,98] and the adiabatic entropy production with the house keeping heat [92,99,100]. That is we can identify contributions to physical heat flows associated with storage and transfer of information.

Next, we turn our attention to the behavior of these contributions to the total entropy production. First, we consider the situation outlined above, that of Markovian bipartite dynamics before considering how they behave when the bipartite assumption is dropped.

VI. BIPARTITE SYSTEMS

In this section we explore the behavior of the entropy productions associated with the storage and transfer of information in systems that are described through the dynamics of two variables X and Y and where the dynamics are of these variables are both bipartite and Markov. To reiterate, these are systems where, if jumps in the variables are permitted, they cannot occur simultaneously in both X and Y. Continuous bipartite dynamics can be considered by considering the

limiting behavior of such systems resulting in a conditional independence condition.

It is important to understand that such dynamics may describe both a system comprising two components of a total physical system in contact with a heat reservoir, or a single variable X controlled by switching protocol Y as the dynamic assumptions in both cases are identical. As such, in this section we first consider the contributions associated with system X in the context of a protocol Y, before describing the joint contributions of the composite system.

A. Single variable and switching protocol

Here, we consider the simplest systems available for consideration. These systems are the typical, canonical, systems found in the original literature of stochastic thermodynamics consisting of a single particle or object of interest denoted the "system" (which may be a composite system itself), represented here by X, and a switching protocol that determines the nature of the dynamics through the change of a potential/nature of a heat bath, etc., represented here by Y. In addition to the specification of Markov and bipartite dynamics, and in contrast to systems we consider subsequently, we do not permit any feedback between the switching protocol and the particle. Consequently, it follows that we may write the probability of observing the joint path $\{x_{[t_0,\tau]}, y_{[t_0,\tau]}\}$,

$$
\begin{aligned}
&p[x_{[t_0,\tau]}, y_{[t_0,\tau]}] \\
&= p(x_{t_0}, y_{t_0}) p[x_{(t_0,\tau]} | x_{t_0}, \{y_{[t_0,\tau)}\}] p[y_{(t_0,\tau]} | y_{t_0}, \{x_{[t_0,\tau)}\}] \\
&= p[x_{(t_0,\tau]} | x_{t_0}, \{y_{(t_0,\tau]}\}] p[y_{(t_0,\tau]} | y_{t_0}] p(x_{t_0} | y_{t_0}) p(y_{t_0}) \\
&= p[x_{[t_0,\tau]} | y_{[t_0,\tau]}] p[y_{[t_0,\tau]}], \tag{40}
\end{aligned}
$$

noting that here the evolution of Y is independent of X, i.e., $p[y_{(t_0,\tau]} | y_{t_0}, \{x_{[t_0,\tau)}\}] = p[y_{(t_0,\tau]} | y_{t_0}]$, and thus that $p[x_{(t_0,\tau]} | x_{t_0}, \{y_{(t_0,\tau]}\}] = p[x_{(t_0,\tau]} | x_{t_0}, y_{[t_0,\tau]}]$.

The division in Eq. (39) identifies heat flows based on information theoretic constructs, on the level of individual behavior, provided by the physical dynamics, but also by the statistics of the ensemble which serves to change the character of the predictability of the process. Explicitly this means changing the properties of the ensemble amounts to changing the amount of information that can be deemed to be in active storage or being transferred and thus the irreversibility associated with it.

As such it is illustrative to consider different situations where one observes different storage and transfer behavior and thus different contributions to the total entropy production. To do so under the system/protocol paradigm without feedback we consider three cases, each with increasing constraint.

1. Irreversible protocols

First, we consider the case where the protocol is chosen randomly according to $p[y_{[t_0,\tau]}]$, and not necessarily reversibly, i.e., $p[y_{[t_0,\tau]}] \neq p^\dagger[y_{[\tau,t_0]}]$. In this case we have, in addition to $\langle \Delta S^x_{\text{storage}} \rangle \geqslant 0$ [which emerges from Eq. (21)], an integral fluctuation theorem regarding the total entropy production in X, plus a change in correlation between system and protocol over the process (captured by a change in a local mutual information, Δi_M, where $\langle i_M \rangle = I_M$ and

$i_M = \ln\{p(x, y)/[p(x)p(y)]\}$). Specifically, we may write

$$\left\langle \exp\left[-\Delta\mathcal{S}_{\text{storage}}^x - \Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y} + \Delta i_M\right]\right\rangle$$

$$= \int dx_{[t_0,\tau]} \int dy_{[t_0,\tau]}$$

$$\times\ p[y_{[t_0,\tau]}]p[x_{(t_0,\tau]}|x_{t_0}, \{y_{[t_0,\tau)}\}]p(x_{t_0}|y_{t_0})$$

$$\times\ \frac{p^{\dagger}[x_{(\tau,t_0]}|x_{\tau}, \{y_{[\tau,t_0)}\}]p(x_{\tau})p(x_{t_0})p(x_{\tau}|y_{\tau})}{p[x_{(t_0,\tau]}|x_{t_0}, \{y_{[t_0,\tau)}\}]p(x_{t_0})p(x_{t_0}|y_{t_0})p(x_{\tau})}$$

$$= \int dy_{[t_0,\tau]}\ p[y_{[t_0,\tau]}]$$

$$\times \int dx_{[t_0,\tau]}p^{\dagger}[x_{(\tau,t_0]}|x_{\tau}, \{y_{[\tau,t_0)}\}]p(x_{\tau}|y_{\tau})$$

$$= 1, \tag{41}$$

such that we have $\langle\Delta\mathcal{S}_{\text{storage}}^x + \Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\rangle - \Delta I_M \geqslant 0$ following from Jensen's inequality. Consequently, we see that we have the bound

$$\left\langle \Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\right\rangle \geqslant \Delta I_M - \left\langle \Delta\mathcal{S}_{\text{storage}}^x\right\rangle \tag{42}$$

or rather

$$\Delta S_x - \Delta I_M + \beta\left\langle\Delta Q_{\text{storage}}^x\right\rangle + \beta\left\langle\Delta Q_{\text{transfer}}^{x \leftarrow y}\right\rangle \geqslant 0, \tag{43}$$

where $\Delta S_x = \langle\Delta\mathcal{S}_{\text{sys}}^x\rangle$. We note that the final mutual information is not strictly necessary for the relation Eq. (43) to hold, meaning that for an initially uncorrelated system and protocol,

we find

$$\left\langle\Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\right\rangle \geqslant -\left\langle\Delta\mathcal{S}_{\text{storage}}^x\right\rangle. \tag{44}$$

Since $\langle\Delta\mathcal{S}_{\text{storage}}^x\rangle \geqslant 0$, $\langle\Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\rangle$ *can* be negative, but not enough so that the total entropy production becomes negative. This forms a central behavior of the two contributions to entropy production: the intrinsic entropy production due to storage is nonnegative, yet the extrinsic entropy production due to transfer may take negative values, potentially offsetting contributions from its surroundings. The above result points out that even when the total entropy production (their sum) is nonnegative the entropy due to transfer may still be negative.

2. Reversible protocols

To see why $\langle\Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\rangle$ can take negative values, it is instructive to consider progressively more restrictive processes. As such we consider a process where there is no irreversibility in the protocol: this can be considered to mean that the evolution of the protocol is purely diffusive, in both the forward and reverse processes, or rather that the protocol in the reverse process is guaranteed to be the time-reversed protocol of the forward process. Alternatively, one may simply posit that protocols are chosen from a distribution that is symmetric with respect to time reversal. All, however, yield the same condition, $p[y_{[t_0,\tau]}] = p^{\dagger}[y_{[\tau,t_0]}]$. When such a condition holds, in addition to the positivity of the storage entropy production $\langle\Delta\mathcal{S}_{\text{storage}}^x\rangle \geqslant 0$, $\langle\Delta\mathcal{S}_{\text{sys}}^x + \beta\Delta Q_{\text{storage}}^x\rangle \geqslant 0$, following from Eq. (21), we may write

$$\left\langle \exp\left[\Delta i_M - \Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\right]\right\rangle = \int dx_{[t_0,\tau]} \int dy_{[t_0,\tau]} p(x_{t_0}|y_{t_0})p[x_{(t_0,\tau]}|x_{t_0}, \{y_{[t_0,\tau)}\}]p[y_{[t_0,\tau]}]$$

$$\times\ \frac{p(x_{t_0})}{p(x_{\tau})}\frac{p(x_{\tau}|y_{\tau})}{p(x_{t_0}|y_{t_0})}\frac{p[x_{(t_0,\tau]}|x_{t_0}]}{p[x_{(t_0,\tau]}|x_{t_0}, \{y_{[t_0,\tau)}\}]}\frac{p^{\dagger}[x_{(\tau,t_0]}|x_{\tau}, \{y_{[\tau,t_0)}\}]}{p^{\dagger}[x_{(\tau,t_0]}|x_{\tau}]}$$

$$= \int dx_{[t_0,\tau]} \int dy_{[t_0,\tau]}\ p(x_{\tau}|y_{\tau})p^{\dagger}[x_{(\tau,t_0]}|x_{\tau}, \{y_{(\tau,t_0)}\}]p[y_{[t_0,\tau]}]\frac{p[x_{[t_0,\tau]}]}{p^{\dagger}[x_{[\tau,t_0]}]}$$

$$= \int dx_{[t_0,\tau]} \int dy_{[t_0,\tau]}\ p(x_{\tau}|y_{\tau})p^{\dagger}[x_{(t_0,\tau]}|x_{\tau}, \{y_{[t_0,\tau)}\}]p^{\dagger}[y_{[\tau,t_0]}]\frac{p[x_{[t_0,\tau]}]}{p^{\dagger}[x_{[\tau,t_0]}]}$$

$$= \int dx_{[t_0,\tau]}\ p^{\dagger}[x_{[\tau,t_0]}]\frac{p[x_{[t_0,\tau]}]}{p^{\dagger}[x_{[\tau,t_0]}]} = 1, \tag{45}$$

such that, again, by Jensen's inequality,

$$-\Delta I_M + \left\langle\Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\right\rangle = -\Delta I_M + \left\langle\beta\Delta Q_{\text{transfer}}^{x \leftarrow y}\right\rangle \geqslant 0. \tag{46}$$

That is, in this picture the single relation of Eq. (43) decomposes into two distinct relations that we may write as two Clausius-like statements for these processes, namely,

$$\Delta S_x \geqslant -\beta\int_{t_0}^t \left\langle dQ_{\text{storage}}^x\right\rangle,$$

$$-\Delta I_M \geqslant -\beta\int_{t_0}^t \left\langle dQ_{\text{transfer}}^{x \leftarrow y}\right\rangle, \tag{47}$$

differences in uncertainty and correlation, bound heat flows associated with storage and transfer of information, respec-

tively. That is, both reflect heat flows that are associated with behavior attributable to X and its interaction with extrinsic variables. Crucially, however, we must understand that the absence of irreversibility in the extrinsic process has lead to the decomposition into two separate bounds facilitated by the stronger bound on $\langle\Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y}\rangle$. In other words, there is no irreversibility extrinsic to X which the entropy due to information transfer need offset and thus the additional bound emerges.

3. Deterministic protocols

Finally, for completeness, we can consider the more traditional conception of a process driven by a switching protocol, namely, that there is a single protocol

$y_{[t_0,\tau]} = \overrightarrow{y}$, which can be reliably selected, repeatedly, such that we have $p[y_{[t_0,\tau]}] = \delta(y_{[t_0,\tau]} - \overrightarrow{y})$. This in turn leads to $p[x_{(t_0,\tau]}|x_{t_0}] \rightarrow p[x_{(t_0,\tau]}|x_{t_0}, \{\overrightarrow{y}\}]$ and thus $\Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow y} = \Delta\mathcal{S}_{\text{transfer}}^{x \leftarrow \emptyset} = 0$ for all realizations. This simply serves to illustrate the fact that information is only deemed to be in transfer when there is a nonsingular ensemble associated with the extrinsic behavior, such that the division of entropy into stored and transferred components is trivial here.

In summary, we see that when there is no variation in the extrinsic process, no transfer of information occurs and thus no entropy production can be associated with transfer of information since knowledge of the protocol adds no predictive capacity to the underlying computations. When this condition is relaxed, but we assert that the extrinsic process is perfectly reversible, the entropy due to transfer of information behaves like an additional physical heat flow, leading to two distinct second laws, as it is rigorously bounded by a change in uncertainty in the process. Finally, we see that this second bound is lost when the dynamics of the protocol can be intrinsically irreversible.

The mechanism by which this happens is important. When the protocol is irreversible, on average the realized $y_{[t_0,\tau]}$ is more probable than the corresponding reverse protocol, $y_{[t_0,\tau]}^{\dagger} = y_{[\tau,t_0]}$. By definition, the forward protocol is a less surprising event in the ensemble that characterizes the marginalized dynamics $p[x_{(t_0,\tau]}|x_{t_0}]$ than the reverse protocol. This indicates that the reverse protocol has more predictive power in the reverse process because it is *rarer* than the forward protocol. This in turn leads to a reduced entropy production arising from the transfer of information: one gains more predictive power using Y in the reverse process than in the forward process *precisely because it is rarer*. This property, where the entropy production due to transfer of information is (negatively) sensitive to the irreversibility of the extrinsic processes is a feature that underlies much of its behavior in more complicated processes.

Finally, we note in this section that we have introduced boundary mutual information terms, required for the relevant integral fluctuation theorems to hold, in a seemingly *ad hoc* and unprincipled way. Indeed, similar terms appear in the literature [101]. In the next section we offer a rationalization for the inclusion of such terms based on a principle of conservation of predictive capacity.

B. Bipartite systems with feedback

Next, we extend the systems considered above to systems where X and Y, while remaining bipartite, may evolve dependently on each other such that they effect feedback. This allows one to consider models of protocols (Y) which can respond to the system (X) or to consider a joint, composite, system, $\{X, Y\}$, identifying contributions to the entropy production from each.

Especially relevant to the latter case is the fact that we may observe the evolution on different levels of description, each of which on has distinct characterizations both in terms of their thermodynamics and their intrinsic computation. So far we have only discussed the intrinsic computation and thermodynamics of individual subsystems, but to understand the role of entropy exchange in terms of intrinsic computation

in such systems we must also understand those behaviours on the level of the composite system.

To do so we must discuss the intrinsic computation, computational irreversibility and thus entropy production associated with the joint system. The joint system is one with only an intrinsic component and thus has a predictive capacity equal to its information storage; i.e.,

$$c_{xy} = a_{xy}. \tag{48}$$

We now seek to relate this predictive capacity to the information processing that is occurring on the level of the individual subsystems characterized through c_x and c_y. To do so we posit a principle that any coherent characterization of the composite system should assign predictive capacity to subcomponents in such a way so as to conserve the total of the joint system. To this end we write

$$c_{xy} = c_x + c_y + c_{\text{int}}^{xy}, \tag{49}$$

where

$$c_x = a_x + t_{y \rightarrow x}, \quad c_y = a_y + t_{x \rightarrow y} \tag{50}$$

are the predictive capacities of the individual subsystems. We emphasize the appearance of c_{int}^{xy}, a new quantity, which we call the the *interaction prediction*. This quantifies the predictive power that emerges only when jointly predicting both X and Y, i.e. the portion of the computational signature of the joint process which cannot be uniquely attached to any individual sub-system and arises from this necessity of conserving predictive capacity. Importantly, unlike c_{xy}, c_x, c_y or their constituent quantities, c_{int}^{xy} has, in general, no bound on its sign in expectation *except* for bipartite systems where it must be negative. This occurs for bipartite systems since in this case c_{int}^{xy} is given by the negative pointwise mutual information,

$$c_{\text{int}}^{xy} = i_M = -\ln\frac{p(x,y)}{p(x)p(y)}. \tag{51}$$

Further, to relate such a decomposition to the computational irreversibilities and entropy production contributions we must attribute an irreversibility term to this interaction prediction. Using the definition of time-reversed computations laid out in Sec. III we may identify an appropriate *time-reversed interaction prediction*,

$$c_{xy}^{\text{int},\dagger} = c_{xy}^{\dagger} - c_x^{\dagger} - c_y^{\dagger}, \tag{52}$$

which analogously leads to an *interaction irreversibility*,

$$d\mathcal{I}_{xy}^{\text{int}} = c_{xy}^{\text{int}} - c_{xy}^{\text{int},\dagger}. \tag{53}$$

Since in these systems we identify $d\mathcal{I}_{xy} = d\mathcal{S}_{\text{tot}}^{xy}$ and $d\mathcal{I}_x = d\mathcal{S}_{\text{tot}}^x$, we identify the interaction irreversibility with an *interaction entropy production*, $d\mathcal{I}_{xy}^{\text{int}} = d\mathcal{S}_{\text{int}}^{xy}$. Importantly, for these bipartite structures this simply reduces to the negative change in pointwise mutual information,

$$\Delta\mathcal{S}_{\text{int}}^{xy} = -\Delta i_M, \quad \langle\Delta\mathcal{S}_{\text{int}}^{xy}\rangle = -\Delta I_M. \tag{54}$$

This then provides a coherent framework for the inclusion of the mutual information in Eqs. (43) and (47): it is accounting for computational irreversibility which cannot be associated with either X or Y. Moreover, we shall see that such a principle

forms the basis of a general approach to systems where fewer assumptions about the dynamics hold.

Proceeding, as mentioned, a particular benefit of the bipartite assumption is that it explicitly allows us to establish heat flows associated with each component of the system, since the conditional independence of the transition probabilities leads to a joint local detailed balance condition decomposing into separate contributions from both X and Y [15] such that we can write

$$
\begin{aligned}
\Delta \mathcal{S}_{\text{tot}}^{xy} &= \Delta \mathcal{S}_{\text{int}}^{xy} + \Delta \mathcal{S}_{\text{tot}}^{x} + \Delta \mathcal{S}_{\text{tot}}^{y} \\
&= -\Delta I_M + \Delta S_{\text{storage}}^{x} + \Delta S_{\text{transfer}}^{x \leftarrow y} \\
&\quad + \Delta S_{\text{storage}}^{y} + \Delta S_{\text{transfer}}^{y \leftarrow x} \\
&= -\Delta I_M + \Delta S_{\text{sys}}^{x} + \beta \Delta Q_{\text{storage}}^{x} + \beta \Delta Q_{\text{transfer}}^{x \leftarrow y} \\
&\quad + \Delta S_{\text{sys}}^{y} + \beta \Delta Q_{\text{storage}}^{y} + \beta \Delta Q_{\text{transfer}}^{y \leftarrow x},
\end{aligned} \tag{55}
$$

with the division of heat flows obeying

$$
\begin{aligned}
\Delta Q^{xy} &= \Delta Q^{x} + \Delta Q^{y} \\
&= \Delta Q_{\text{storage}}^{x} + \Delta Q_{\text{transfer}}^{x \leftarrow y} + \Delta Q_{\text{storage}}^{y} + \Delta Q_{\text{transfer}}^{y \leftarrow x}.
\end{aligned} \tag{56}
$$

Importantly, however, no individual contribution, e.g. $\Delta S_{\text{tot}}^{x} = \Delta S_{\text{storage}}^{x} + \Delta S_{\text{transfer}}^{x \leftarrow y}$, nor, even sums of contributions, $\Delta S_{\text{tot}}^{x} + \Delta S_{\text{tot}}^{y}$, has a bound on its sign in expectation; in the context of multi-component systems the irreversibility of one component can be negative if it is commensurately offset by positive irreversibility in the remaining components. The second law applies to the total system (and environment), not necessarily on each component in isolation. Importantly, however, the remaining components which offset such a reduction *must include the interaction entropy production.*

The precise nature of how this phenomenon manifests has attracted much attention in terms of information flows, having its origins in measurement and feedback systems, identifying minimal correlations that can be exploited to reduce the entropy of a subsystem. In this work we take a different approach and identify the *physical entropy* that can be offset in different subsystems, arguing that it is the entropy production due to information transfer that can cause negative entropy productions and indeed that it is the entropy production due to information transfer associated with other subsystems that have the ability to offset this reduction. The bounds that then arise from the these interactions then work in interplay with the interaction entropy production.

First, we establish that the total entropy of the entire system is, necessarily, positive in expectation in accordance with the second law, formulated here as a contribution entirely associated with information storage and thus is rigorously nonnegative by Eq. (21), i.e.,

$$
\begin{aligned}
\langle \Delta \mathcal{S}_{\text{tot}}^{xy} \rangle &= \int dx_{[t_0, \tau]} \int dy_{[t_0, \tau]} \, p[x_{[t_0, \tau]}, y_{[t_0, \tau]}] \\
&\quad \times \ln \frac{p[x_{[t_0, \tau]}, y_{[t_0, \tau]}]}{p^{\dagger}[x_{[\tau, t_0]}, y_{[\tau, t_0]}]} \geqslant 0.
\end{aligned} \tag{57}
$$

Next, however, we consider the following quantity

$$
\begin{aligned}
&\langle \exp \left[-(\Delta \mathcal{S}_{\text{int}}^{xy} + \Delta \mathcal{S}_{\text{tot}}^{x} + \Delta \mathcal{S}_{\text{tot}}^{y} - \Delta \mathcal{S}_{\text{storage}}^{x}) \right] \rangle \\
&= \int dx_{[t_0, \tau]} \int dy_{[t_0, \tau]} \, p(x_{t_0}, y_{t_0}) p[x_{(t_0, \tau]}, y_{(t_0, \tau]} | x_{t_0}, y_{t_0}] \\
&\quad \times \frac{p(x_\tau, y_\tau) p^{\dagger}[x_{(\tau, t_0]}, y_{(\tau, t_0]} | x_\tau, y_\tau]}{p(x_{t_0}, y_{t_0}) p[x_{(t_0, \tau]}, y_{(t_0, \tau]} | x_{t_0}, y_{t_0}]} \frac{p[x_{[t_0, \tau]}]}{p^{\dagger}[x_{[\tau, t_0]}]} \\
&= \int dx_{[t_0, \tau]} \, p^{\dagger}[x_{[\tau, t_0]}] \frac{p[x_{[t_0, \tau]}]}{p^{\dagger}[x_{[\tau, t_0]}]} \\
&= 1, \tag{58}
\end{aligned}
$$

such that we have the bound

$$
\langle \Delta \mathcal{S}_{\text{tot}}^{xy} - \Delta \mathcal{S}_{\text{storage}}^{x} \rangle \geqslant 0, \tag{59}
$$

which may be considered equivalent to

$$
\langle \Delta \mathcal{S}_{\text{int}}^{xy} \rangle + \langle \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y} \rangle \geqslant -\langle \Delta \mathcal{S}_{\text{tot}}^{y} \rangle. \tag{60}
$$

This, along with the property $\langle \Delta \mathcal{S}_{\text{storage}}^{y} \rangle \geqslant 0$ indicates that if the total entropy production of subsystem Y is negative, it arises because of *negative* $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle$ components that outweigh the information storage entropy, but moreover that this negative entropy is balanced by the entropy due to *transfer* of information into X along with the interaction entropy. In the bipartite systems considered here, in the steady state, this may be expressed

$$
\langle \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y} \rangle \geqslant -\langle \Delta \mathcal{S}_{\text{tot}}^{y} \rangle \tag{61}
$$

since the mutual information is unchanged, demonstrating explicitly how the extrinsic entropy production in X controls the second law breakage in Y.

Again, as with the treatment of switching protocols without feedback, we see a direct inverse relationship between the irreversibility in an extraneous time series, here Y, and the irreversibility in X that arises from information transfer from that variable. Despite the presence of feedback, the mechanism is the same as in the case of a stochastic protocol. With increased irreversibility in Y, reflected in a positive $\Delta \mathcal{S}_{\text{tot}}^{y}$, most sequences $y_{[t_0, \tau]}$ are common trajectories and thus well reflected in the coarse grained dynamics making them not particularly informative, while $y_{[t_0, \tau]}^{\dagger} = y_{[\tau, t_0]}$ are rare and are thus more informative leading to negative entropy production due to information transfer. However, when $\Delta \mathcal{S}_{\text{tot}}^{y}$ is negative, sequences $y_{[t_0, \tau]}$ are chosen that are relatively uncommon (with respect to their time reverse, $y_{[t_0, \tau]}^{\dagger} = y_{[\tau, t_0]}$, in construction of the marginal dynamics), leading to more informative realizations in the dynamics of Y in the forward time direction and thus a positive entropy production due to information transfer.

1. Recovery and generalization of previous information-thermodynamic bounds

Here we relate the bounds that we have derived, based on describing the thermodynamic entropy production as formed from information processing quantities, with previous results in the literature which generally seek to sharpen or generalize the second law with information processing quantities.

A well-known result concerning the transfer entropy rate in the steady state found in [16,102,103] reads

$$\langle \Delta \dot{\mathcal{S}}_{\text{tot}}^x \rangle + \dot{T}_{x \to y} \geqslant 0, \tag{62}$$

which may be integrated over finite times by considering the pathwise transfer entropy [82],

$$\mathcal{T}_{x \to y}\big[x_{[t_0,t]}, y_{[t_0,t]}\big] = \ln \frac{p\big[y_{(t_0,t]}\big|y_{t_0}, \{x_{[t_0,t)}\}\big]}{p\big[y_{(t_0,t]}\big|y_{t_0}\big]}, \tag{63}$$

such that

$$\langle \Delta \mathcal{S}_{\text{tot}}^x \rangle + \langle \mathcal{T}_{x \to y} \rangle \geqslant 0. \tag{64}$$

Outside of the steady state an analogous expression for Bayesian networks was given in Ref. [101] by considering the addition of a mutual information term, which after adapting to continuous time processes yields

$$\langle \Delta \mathcal{S}_{\text{tot}}^x \rangle + \langle \mathcal{T}_{x \to y} \rangle - \Delta I_M \geqslant 0. \tag{65}$$

We emphasize, the distinction between this relation and the analogous relation in our framework,

$$\langle \Delta \mathcal{S}_{\text{tot}}^x \rangle + \langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle + \langle \Delta \mathcal{S}_{\text{int}}^{xy} \rangle \geqslant 0, \tag{66}$$

is that all terms in the latter are understood as components of the total entropy production of the composite system by considering the decomposition of the observed irreversibilities along computational lines. In this way, $\langle \Delta \mathcal{S}_{\text{int}}^{xy} \rangle = -\Delta I_M$ is not an *ad hoc* introduction to satisfy the bound, but the irreversibility associated with the joint prediction and computation in the system which must be accounted for. In contrast, $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle \neq \langle \mathcal{T}_{x \to y} \rangle$, with the former identified as a component of the entropy production in Y while the latter is a pure information term. However, our formalism can be seen to reduce to the above if a specific choice of reverse process is considered, common to the original literature on feedback protocols [21], whereby the reverse process forces the extraneous variable, e.g., Y, to exactly retrace the path it traced in the forward realization, independently of X such that the reverse dynamics are distinct from the forward dynamics, i.e., $p \neq p^\dagger$. This has the specific result of setting $p^\dagger[y_{[\tau,t_0]}, \{x_{[\tau,t_0]}\}] = p[y_{[t_0,\tau]}] = p^\dagger[y_{[\tau,t_0]}]$, precisely returning $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle = \langle \mathcal{T}_{x \to y} \rangle$.

This may give some insight into the thermodynamic meaning, if any, of the bare transfer entropy as it is equal to the *irreversibility* $\langle \Delta \mathcal{I}_{y \leftarrow x}^T \rangle$ analogous to the the entropy production $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle$ in the case of nonautonomous feedback. That is, the transfer entropy is a limit case of $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle$ when Y has no direct energetic, or thermal, interpretation as in the case of a nonautonomous feedback controller.

There is no general bound characterizing the relationship between $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle$ and $\langle \mathcal{T}_{x \to y} \rangle$, with the former in certain situations being larger or smaller than the transfer entropy depending on the nature of the dynamics and the reverse behavior. However, their equality in the case of nonautonomous protocols, where it is relevant, allows confirmation of the bound in cases where $\langle \mathcal{T}_{x \to y} \rangle$ is computable, but $\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \rangle$ is not, which we shall find of use when we illustrate the generalized application of our results in a nonbipartite system consisting of a thermodynamic system of interest and a feedback controller.

VII. BEYOND THE BIPARTITE ASSUMPTION

In this section we apply our formalism to systems that do not possess the bipartite structure considered thus far. While the thermodynamics and entropy production of such systems may, at least in the case of discrete state spaces, be treated as a whole using conventional stochastic thermodynamics, treatment of the entropy production associated with their individual components ordinarily cannot. As discussed, this inability arises because under such dynamics the heat exported to the environment does not divide into components that can be exclusively associated with each subsystem. Rather, at least in some sense, there must be a component of the heat flow which is shared between the subsystems. When considering only discrete space, continuous time systems such as those described as a master equation it might seem natural to categorize such heat flows based on the nature of the transition, with a shared heat flow being associated with those not permitted in bipartite systems, namely joint transitions. However, in effect, this is merely categorizing distinct transitions as bipartite (individual) or nonbipartite (joint) and leaves fundamental questions unanswered. For instance, if *all* transitions are joint transitions one would conclude that all heat flow should be identified as belonging to the joint system. However, one may still track only the probabilistic behavior of X in the context of Y and seek to characterize its irreversibility. Is this irreversibility related to the irreversibility of the joint system or that of subsystem Y? What if only the energetics of one sub-system can be defined if the other corresponds to a nonthermal process such as a feedback controller? These questions go entirely unanswered if the formalism simply assigns all irreversibility to the composite system. Furthermore, in the limit of continuous dynamics the bipartite condition interpreted as noncoincident transitions is no longer applicable with such a property resulting in conditional independence in the dynamics. In these systems both X and Y simultaneously fluctuate on all timescales and so the identification of heat flows is generally nontrivial, but not necessarily ill defined. For instance, as we shall explore later, such continuous, nonbipartite, systems, if representing a physical system and correlated feedback controller, the heat dissipated by the physical system can be captured by the computational irreversibility.

It is in answering these questions that we propose our formalism has particular merit since, because it is based on information theory, it is agnostic to the precise nature of the physical dynamics which are leading to the probability distributions, yet is consistent with conventional stochastic thermodynamics where heat flows can be identified as per the previous section.

Explicitly we concern ourselves with systems where the relation in Eq. (31) does not hold; i.e.,

$$\lim_{dt \to 0} p(x_{t+dt}, y_{t+dt}|x_t, y_t)$$
$$\neq \lim_{dt \to 0} p(x_{t+dt}|x_t, y_t)p(y_{t+dt}|x_t, y_t). \tag{67}$$

To proceed, however, we take precisely the same course of action as before, establishing the interaction irreversibility based on a principle of conservation of predictive capacity according to Eq. (49), which must hold independently of the physical energetics of the system or whether they are

identifiable. This is given by

$$
\begin{aligned}
\Delta \mathcal{I}_{xy}^{\text{int}} = &- \Delta I_M + \ln \frac{p\big[x_{(t_0,\tau]}, y_{(t_0,\tau]} \big| x_{t_0}, y_{t_0}\big]}{p^{\dagger}\big[x_{(\tau,t_0]}, y_{(\tau,t_0]} \big| x_t, y_t\big]} \\
&+ \ln \frac{p^{\dagger}\big[x_{(\tau,t_0]} \big| x_t, \{y_{[\tau,t_0)}\}\big] p^{\dagger}\big[y_{(\tau,t_0]} \big| y_t, \{x_{[\tau,t_0)}\}\big]}{p\big[x_{(t_0,\tau]} \big| x_{t_0}, \{y_{[t_0,\tau)}\}\big] p\big[y_{(t_0,\tau]} \big| y_{t_0}, \{x_{[t_0,\tau)}\}\big]} \\
= &- \Delta I_M + \beta \Delta Q_{\text{int}}^{xy}.
\end{aligned}
\tag{68}
$$

Explicitly, the interaction irreversibility is comprised of a change in mutual information which appeared in the bipartite description *and* a heat contribution that did not appear. Again, this heat is *not* identified by joint transitions in the composite system, but rather by the irreversibility which cannot be observed in the individual subsystems. At this point we remain agnostic as to whether $\Delta \mathcal{S}_{\text{tot}}^x$ has a precise meaning in these circumstances (though we shall argue that it can in a subsequent example) and so strictly we may deal with the computation irreversibilities $\Delta \mathcal{I}_x = \Delta \mathcal{I}_x^A + \Delta \mathcal{I}_{x \leftarrow y}^T$. However, the total entropy production of the composite system is well defined and so we may still identify the contributions to this composite entropy production in terms of storage and transfer, i.e., $\Delta \mathcal{I}_x = \Delta \mathcal{S}_{\text{storage}}^x + \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y}$ and $\Delta \mathcal{I}_{\text{int}}^{xy} = \Delta \mathcal{S}_{\text{int}}^{xy}$, such that

$$
\begin{aligned}
\Delta \mathcal{S}_{\text{tot}}^{xy} = &\, \Delta \mathcal{I}_x + \Delta \mathcal{I}_y + \Delta \mathcal{I}_{xy}^{\text{int}} \\
= &\, \Delta \mathcal{S}_{\text{storage}}^x + \Delta \mathcal{S}_{\text{storage}}^y \\
&+ \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y} + \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} + \Delta \mathcal{S}_{\text{int}}^{xy} \\
= &\, \Delta \mathcal{S}_{\text{sys}}^x + \beta \Delta Q_{\text{storage}}^x + \Delta \mathcal{S}_{\text{sys}}^y + \beta \Delta Q_{\text{storage}}^y \\
&+ \beta \Delta Q_{\text{transfer}}^{x \leftarrow y} + \beta \Delta Q_{\text{transfer}}^{y \leftarrow x} - \Delta I_M + \beta \Delta Q_{\text{int}}^{xy},
\end{aligned}
\tag{69}
$$

such that

$$
\begin{aligned}
\Delta Q^{xy} = &\, \Delta Q_{\text{storage}}^x + \Delta Q_{\text{storage}}^y + \Delta Q_{\text{transfer}}^{x \leftarrow y} \\
&+ \Delta Q_{\text{transfer}}^{y \leftarrow x} + \Delta Q_{\text{int}}^{xy}.
\end{aligned}
\tag{70}
$$

At this point we reiterate: These are divisions of the total heat, which is well defined, but due to the lack of bipartite dynamics we cannot necessarily associate $\Delta Q_{\text{storage}}^x + \Delta Q_{\text{transfer}}^{x \leftarrow y} = \Delta Q^x$ as we could in the previous section.

Importantly, however, these entropy productions follow *exactly* the same relationships as for the bipartite case, by construction, with the same integral fluctuation theorems holding, namely,

$$
\big\langle \exp\big[-\big(\Delta \mathcal{S}_{\text{int}}^{xy} + \Delta \mathcal{S}_{\text{storage}}^x + \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y} + \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x}\big)\big]\big\rangle = 1,
\tag{71}
$$

again such that

$$
\begin{aligned}
&\big\langle \Delta \mathcal{S}_{\text{int}}^{xy} \big\rangle + \big\langle \Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x} \big\rangle \\
&\qquad \geqslant -\big\langle \Delta \mathcal{S}_{\text{storage}}^x \big\rangle - \big\langle \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y} \big\rangle = -\big\langle \Delta \mathcal{I}^x \big\rangle,
\end{aligned}
\tag{72}
$$

as before, lending weight to the generality of the proposed measures of irreversibility.

Moreover, the above bound also holds when $\Delta \mathcal{S}_{\text{transfer}}^{y \leftarrow x}$ is replaced by the bare pathwise transfer entropy $\mathcal{T}_{x \to y}$, as in the bipartite case, allowing us to write

$$
\big\langle \Delta \mathcal{I}_x \big\rangle + \big\langle \Delta \mathcal{I}_{xy}^{\text{int}} \big\rangle \geqslant -\big\langle \mathcal{T}_{x \to y} \big\rangle
\tag{73}
$$

or rather

$$
\big\langle \Delta \mathcal{I}_{xy} \big\rangle - \big\langle \Delta \mathcal{I}_y \big\rangle \geqslant -\big\langle \mathcal{T}_{x \to y} \big\rangle,
\tag{74}
$$

which may equivalently be expressed in the rate form $\langle \Delta \dot{\mathcal{I}}_x \rangle + \langle \Delta \dot{\mathcal{I}}_{xy}^{\text{int}} \rangle \geqslant -\dot{T}_{x \to y}$. This is a central result and can be seen to generalize the well known information-thermodynamic bound $\langle \Delta \dot{\mathcal{S}}_{\text{tot}}^x \rangle \geqslant -\dot{T}_{x \to y}$ and $\langle \Delta \dot{\mathcal{S}}_{\text{tot}}^x \rangle \geqslant \Delta \dot{I}_M - \dot{T}_{x \to y}$ explored in depth in several other works [16,101–104] to the case of nonbipartite dynamics. Explicitly, when bipartite dynamics are employed $\Delta \dot{\mathcal{I}}_x$ reduces to $\Delta \dot{\mathcal{S}}_{\text{tot}}^x$ and $\Delta \dot{\mathcal{I}}_{xy}^{\text{int}}$ reduces to $\Delta \dot{I}_M$, which then vanishes in the steady state. It is important to note that in both regimes the transfer entropy $\dot{T}_{x \to y}$ bounds the irreversibility associated with the subsystem X *and* the irreversibility associated with its interaction with Y.

It is interesting to consider what physical systems might be described by such dynamics that do not permit bipartite form. Two major examples can be considered. The first is when the whole system is well defined thermodynamically, with defined heat flows and dissipation, but has interacting parts that evolve in such a way that it cannot be traditionally decomposed any further. Alternatively, when considering autonomous measurement and feedback one can imagine a feedback mechanism that does not meet the bipartite assumption. In both cases either the total heat of the composite system or the heat flow in the system being monitored should still be identifiable. However, currently no formalism exists for characterizing any such division of this total heat or for characterizing the feedback performance in these situations. A particular case will be treated in the next section.

VIII. EXAMPLES

A. Minimal bipartite model

Here we illustrate some of the behavior of the entropy production contributions with a simple model. This model consists of two subsystems interacting with a bipartite structure. The first subsystem, X, is composed of three states, $x \in \mathcal{X} = \{A, B_1, B_2\}$, while the second subsystem, Y, is composed of two states, $y \in \mathcal{Y} = \{1, 2\}$, such that there are six effective global states, $\{x, y\} \in \mathcal{X} \otimes \mathcal{Y}$. The bipartite structure insists upon strict conditional independence in the updating of these subsystems, which, in continuous time, means that there are no instantaneous joint transitions in both X and Y with probability 1.

The structure is an elaboration on the minimal model introduced in Ref. [103], but with an additional state in subsystem X. Further, we insist upon a particular transition rate structure which entails the transition rates governing subsystem Y only being dependent on whether X is in state A or the collection of states $B = \{B_1, B_2\}$, such that it cannot distinguish between states B_1 and B_2 that comprise B. This structure is illustrated in Fig. 2 along with specific transition rates, specified up to the transition rates between the substates B_1 and B_2. Other transitions are parametrized by the dimensionless constants $\{r_x, r_y\} \in [-1, 1] \otimes [-1, 1]$ and timescale parameters $\gamma_x > 0$, $\gamma_y > 0$.

This structure allows us to separate the mechanisms that pertain to irreversibility associated with storage and transfer of information. This is because the transitions between B_1 and
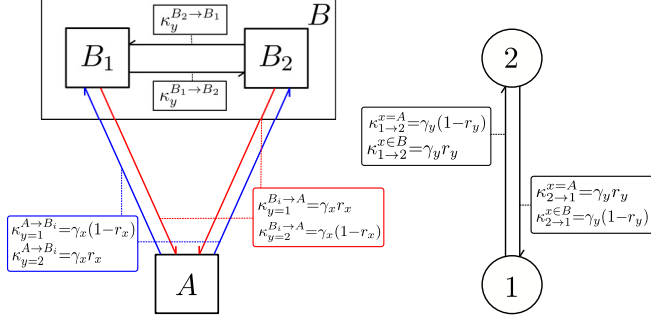
FIG. 2. Coupled three- and two-state systems. System X (shown on left) has transition rates dependent on the instantaneous state of system Y. System Y (shown on right) has transition rates dependent on the instantaneous state of system X. System Y can only depend on the coarse states A and B in X: the transition rates of system Y *do not* distinguish between states B_1 and B_2. Since the system is bipartite, no joint transitions in systems X and Y can occur.

B_2 can introduce irreversibility independently of Y, while the transitions between the pairs A, B and 1, 2 explicitly interact with each other but do not introduce any net steady current into either subsystem when viewed in isolation.

To this end we consider a parametrization of the transition rates $\kappa_y^{B_1 \leftrightarrow B_2}$ as follows:

$$\begin{aligned}
\kappa_y^{B_1 \to B_2} &= \gamma_x k^+, \quad \forall y \in \{1, 2\}, \\
\kappa_y^{B_2 \to B_1} &= \gamma_x k^-, \quad \forall y \in \{1, 2\},
\end{aligned} \tag{75}$$

where k^+ and k^- are arbitrary constants in (0,1). Considering the steady state and solving the master equation by standard methods then yields

$$\begin{aligned}
\langle \Delta \dot{S}_{\text{storage}}^x \rangle &= \frac{\gamma_x(r_x + r_y - 2r_x r_y)(k^+ - k^-)\ln[k^+/k^-]}{3[(k^+ + k^-) + (r_x + r_y - 2r_x r_y)]} \\
&\quad + O(a), \\
\langle \Delta \dot{S}_{\text{transfer}}^{x \leftarrow y} \rangle &= \frac{4}{3}\gamma_x(r_x - r_y)\ln\left[\frac{1 - r_x}{r_x}\right] + O(a),
\end{aligned}$$

$$\begin{aligned}
\langle \Delta \dot{S}_{\text{transfer}}^{y \leftarrow x} \rangle &= \frac{4}{3}\gamma_x(r_y - r_x)\ln\left[\frac{1 - r_y}{r_y}\right] + O(a), \\
\langle \Delta \dot{S}_{\text{storage}}^y \rangle &= 0,
\end{aligned} \tag{76}$$

where $a = \gamma_x/\gamma_y$ such that we have effected a separation of timescales to avoid the difficulty associated with calculation of generally non-Markovian objects. If a further separation of timescales exists, $\kappa_y^{B_1 \to B_2} + \kappa_y^{B_2 \to B_1} = \gamma_x' \ll \gamma_x$ then the storage entropy reduces to $\langle \Delta \dot{S}_{\text{storage}}^x \rangle = (1/3)\gamma_x'(k^+ - k^-)\ln[k^+/k^-] + O(\gamma_x'/\gamma_x)$. The results in Eqs. (76) are illustrated in Fig. 3. Importantly, we see that as we vary the feedback coefficient r_y controlling subsystem Y the entropy due to transfer in each system is offsetting the other, with the entropy production associated with storage only varying weakly based on changes in the stationary solution to the master equation. Indeed, since in this example we have $\langle \Delta \dot{S}_{\text{storage}}^y \rangle = 0$, a stronger bound than Eq. (59) holds, with $\langle \Delta \dot{S}_{\text{transfer}}^{x \leftarrow y} + \Delta \dot{S}_{\text{transfer}}^{y \leftarrow x} \rangle \geqslant 0$ holding completely independently of $\langle \Delta \dot{S}_{\text{storage}}^x \rangle$ in addition to $\langle \Delta \dot{S}_{\text{tot}}^x + \Delta \dot{S}_{\text{transfer}}^{y \leftarrow x} \rangle \geqslant 0$. Moreover, if we consider r_x and r_y to be fixed, varying only the thermodynamic force causing asymmetry in the $B_1 \leftrightarrow B_2$ transition, we see total independence of in the entropy productions due to transfer as the asymmetric transitions are not implicated in the feedback between the two systems. Indeed, we see that the entropy production due to storage in this example manifests exclusively due to stationary current which can only arise when k^+ and k^- are unequal explaining the behavior with varying k^+.

B. Computational and physical irreversibility in nonbipartite autonomous feedback systems

Generally, there are many systems that do not obey the bipartite assumption. These range from composite systems that permit joint transitions between different elements, composite systems where the subsystems experience thermal (or other) noise sources that cannot be reliably idealized as being independent, through to systems traditionally out of scope thermodynamics. However, an investigation of their irreversibility may yet yield analogous bounds important to their behavior.

In this section, however, we consider a particular example, that is a system designed to model autonomous feedback by a
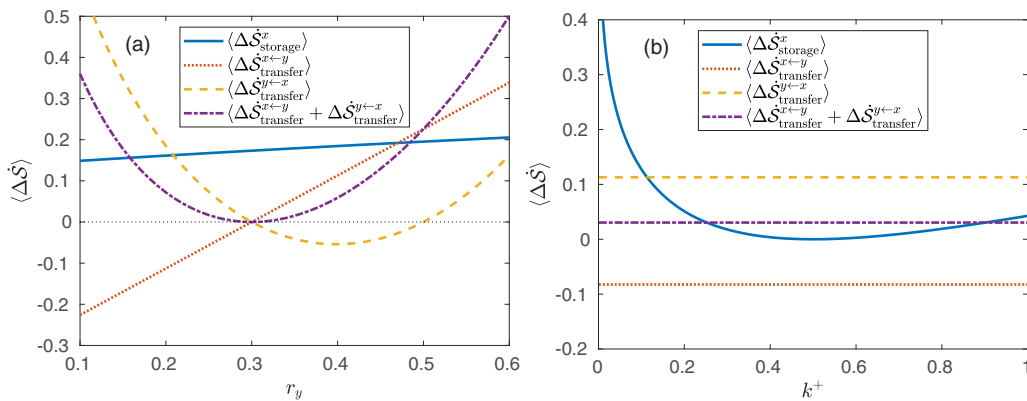

FIG. 3. Entropy productions arising from the coupled three and two state model. In subfigure (a) we have set $\gamma_x = 1$, $k^- = 0.1$, $k^+ = 0.9$, $r_x = 0.3$ and allow r_y to vary. In subfigure (b) we have set $\gamma_x = 1$, $k^- = 0.5$, $r_x = 0.3$, $r_y = 0.3$ and allow k^+ to vary. Both consider zeroth order contributions in $a = \gamma_x/\gamma_y$.

controller, Y, on a physical system of interest X. A minimal mathematical model that achieves this is a pair of coupled Ornstein-Uhlenbeck processes described by the stochastic differential equations

$$dx = Ax\,dt + By\,dt + V_x\,dW_x,$$
$$dy = Cx\,dt + Dy\,dt + V_y\,dW_y, \tag{77}$$

where dW_x and dW_y represent thermal noise and the inherent error in the measurement process, respectively. Normally, these two noise sources are deemed to be statistically independent resulting in bipartite dynamics, the behavior of which has been previously studied [16,102,105]. However, we can quite easily posit that the noise sources *are* correlated for any number of plausible physical reasons, for example noise in the measurement apparatus may be arising, in part, from the same thermal source such that the measurement suffers from noise contamination. In this case we have $\langle dW_x dW_y \rangle = \rho\,dt$ where $\rho \in [-1, 1]$ characterizes the noise correlation such that for $\rho = 1$ the noise experienced by each of X and Y are identical.

This system is governed by a Fokker-Planck equation of the form

$$\partial_t p(x, y, t) = -\boldsymbol{\nabla} \cdot (J_x, J_y),$$
$$J_x = (Ax + By)p(x, y) - \frac{V_x^2}{2} \frac{\partial p(x, y)}{\partial x}$$
$$\qquad - \frac{\rho V_x V_y}{2} \frac{\partial p(x, y)}{\partial y},$$
$$J_y = (Cx + Dy)p(x, y) - \frac{V_y^2}{2} \frac{\partial p(x, y)}{\partial y}$$
$$\qquad - \frac{\rho V_x V_y}{2} \frac{\partial p(x, y)}{\partial x}, \tag{78}$$

which has a Gaussian steady state given in Appendix B.

A well-known bound that arises in information thermodynamics is [16,102,103], as usually stated in the steady state,

$$\frac{d\beta \langle \Delta Q^x \rangle}{dt} + \dot{T}_{x \to y} \geqslant 0, \tag{79}$$

where $\dot{T}_{x \to y}$ is transfer entropy rate. We emphasize, with correlated noise this relation *does not* hold.

1. Identification of information and heat flows

While the framework outlined in Sec. VII describes how to divide entropy productions along computational divisions in terms of predictive capacity to arbitrary systems where individual currents and heats may not be well defined, we emphasize that for continuous systems of the form in Eq. (77) we *can* identify such quantities, under the right circumstances, despite the lack of a bipartite structure.

To do so we point out that the sampling paths are absolutely continuous such that, unlike with master equations with joint transitions, the variables never change discontinuously. As such the notion of a dissipated heat, by subsystem X, for instance, is not inappropriate, but more challenging to identify and depends on the exact nature of the model. For instance, in the present model it is of consequence that X is the thermodynamic system in question, while Y represents a feedback controller. As such, we are not concerned with the

thermodynamics of Y or even the joint system $\{X, Y\}$ and this changes the interpretation. As such, the entropy production of the dynamics of the joint particle and controller system is immaterial, but their computational irreversibilities still are well defined.

To proceed and identify such heat flows associated with the single subsystem X, we must be more careful than just assuming a conventional local detailed balance relation but rather appeal to the arguments in the spirit of the foundational stochastic energetics of Sekimoto [106–108]. First, we consider systems, such as over-damped Langevin dynamics where the energetics are all determined by the random variables such that all differentials are stochastic too. In such situations we state that there is some Hamiltonian, H, controlling X, which is parametrized by a protocol Y. When this protocol is deterministic, this traditionally leads to a stochastic version of the first law as follows:

$$dH(x, y) = \partial_x H(x, y) \circ dx + \partial_y H(x, y)\dot{y}\,dt, \tag{80}$$

identifying changes in internal energy, heat, and work ($dU = dQ + dW$) with the crucial use of Stratonovich integration rules (indicated by the $\circ$ notation), which render the stochastic calculus in the form of the usual thermodynamic differentials. If we allow the protocol Y to be stochastic, we instead write

$$dH(x, y) = \partial_x H(x, y) \circ dx + \partial_y H(x, y) \circ dy, \tag{81}$$

with the division of internal energy into heat and work operating in exactly the same way. Moreover, Eq. (81) remains unchanged under correlated noise such that the heat and work contributions are still identifiable where we note that Stratonovich integration includes correlation terms in the nonbipartite case such that

$$f(x, y) \circ dx = f(x, y)dx + \frac{V_x^2}{2} \frac{\partial f(x, y)}{\partial x}dt$$
$$\qquad + \frac{\rho V_x V_y}{2} \frac{\partial f(x, y)}{\partial y}dt, \tag{82}$$

where absence of the $\circ$ notation indicates the nonanticipating Itō integral. Under this interpretation we might posit some over-damped dynamics such that X is a spatial coordinate, $V_x = \sqrt{2k_B T / m\gamma}$ and $Ax + By = -(m\gamma)^{-1}\partial_x H_x(x, y)$, such that we have a Hamiltonian $H_x(x, y) = -m\gamma[(A + B)x^2 - B(x - y)^2]$, i.e., a harmonic trap with a protocol controlling an additional harmonic term. Importantly, in this system the computational irreversibility reflects the ratio of heat flow to the environment to the environmental temperature, $\beta \partial_x H(x, y) \circ dx$, such that $\Delta \mathcal{I}_x = \Delta \mathcal{S}_{\text{storage}}^x + \Delta \mathcal{S}_{\text{transfer}}^{x \leftarrow y} = \Delta \mathcal{S}_{\text{sys}}^x + \beta \Delta Q^x$.

However, in situations where the system contains degrees of freedom which are not stochastic, but appear in the Hamiltonian an additional correlation term appears in the computational irreversibility which has no thermodynamic analog. In these cases we must instead identify

$$\Delta \mathcal{I}_x = \Delta \mathcal{S}_{\text{sys}}^x + \beta \Delta Q^x + \Delta \mathcal{S}_{\text{corr}}^x, \tag{83}$$

where $\Delta \mathcal{S}_{\text{corr}}^x$ is the difference in heat flows between that which can be physically associated with X and that which is

associated with the irreversibility of X such that

$$d\mathcal{S}_{\text{corr}}^x = \beta dQ_{\text{corr}}^x = \ln \frac{p^\dagger(x_t|x_{t+dt}, y_{t+dt})}{p^\dagger(x_t|x_{t+dt}, y_t)}. \quad (84)$$

This situation arises due to parity differences in the variables separating the current into reversible and irreversible components [109]. Consequently, differential forms for the environmental entropy production introduce a correlation term that would arise from stochastic integration of the reversible current but does not emerge in the stochastic energetics since the energetic term arises from integration with respect to some deterministic variable (i.e., one not subject to noise), e.g., the particle position in a full under-damped Langevin description. We emphasize, in such situations great care must be taken to identify the exact nature of the dynamics and any Hamiltonian, if any, they are emerging from to correctly identify irreversible and reversible currents. We also mention that correlated noise must be treated carefully in systems with odd variables where the noise itself may be required to transform under time reversal [110].

For the remainder we shall consider the case of the former, where, in the steady state we consider $\langle\Delta\mathcal{S}_{\text{storage}}^x + \Delta\mathcal{S}_{\text{transfer}}^{x\leftarrow y}\rangle = \beta\langle\Delta Q^x\rangle$, for brevity, but acknowledge the possibility of this additional step in the identification of the environmental entropy production in different models.

Returning to the model in question we reiterate that the naive application of the traditional information theoretic bound on the computational and physical irreversibility does not hold, i.e.,

$$\frac{d\langle\Delta\mathcal{S}_{\text{storage}}^x\rangle}{dt} + \frac{d\langle\Delta\mathcal{S}_{\text{transfer}}^{x\leftarrow y}\rangle}{dt} + \dot{T}_{x\rightarrow y} \not\geq 0. \quad (85)$$

Instead, to find reliable bounds we must turn to our formulation in terms of computational irreversibilities, which accounts for the interaction entropy production. As such the appropriate bound is

$$\frac{d\langle\Delta\mathcal{S}_{\text{tot}}^x\rangle}{dt} + \frac{d\langle\Delta\mathcal{I}_{xy}^{\text{int}}\rangle}{dt} + \frac{d\langle\Delta\mathcal{S}_{\text{transfer}}^{y\leftarrow x}\rangle}{dt} \geq 0 \quad (86)$$

or

$$\frac{d\langle\Delta\mathcal{S}_{\text{tot}}^x\rangle}{dt} + \frac{d\langle\Delta\mathcal{I}_{xy}^{\text{int}}\rangle}{dt} + \dot{T}_{x\rightarrow y} \geq 0. \quad (87)$$

to account for the nonbipartite dynamics. We note the inclusion of the more general $\Delta\mathcal{I}_{xy}^{\text{int}}$ rather that $\Delta\mathcal{S}_{\text{int}}^{xy}$ in the bound because, as discussed, while the computational irreversibility of the joint particle and controller system is well defined, it is not related to an entropy production in this case.

For the present example we are in a position to illustrate such a bound mathematically. In the steady state, the transfer entropy of this system permits an analytical solution, adapted from Ref. [16] (see also Ref. [111]), derived in Appendix C,

$$\dot{T}_{x\rightarrow y} = \frac{1}{2}\left(\sqrt{A^2 - 2\rho AC\frac{V_x}{V_y} + C^2\frac{V_x^2}{V_y^2}} - \sqrt{A^2} - \rho C\frac{V_x}{V_y}\right). \quad (88)$$

Meanwhile, the rate of computational irreversibility in X in the steady state, here equal to the entropy production rate, is
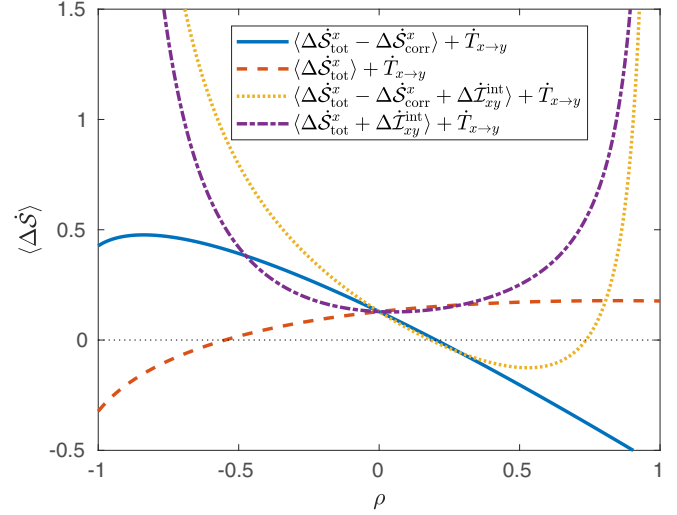

FIG. 4. Coupled linear feedback model. We set $A = -2$, $D = -4$, $B = C = 1$, $V_x = 2$, $V_y = 3/2$, and we allow the correlation, ρ, to vary.

given by

$$\frac{d\langle\Delta\mathcal{S}_{\text{tot}}^x\rangle}{dt} = \frac{B\{CV_x^2 - V_y[\rho(A-D)V_x + BV_y]\}}{(A+D)V_x^2}. \quad (89)$$

For completeness we also point out that for this system we have

$$\frac{d\langle\Delta\mathcal{S}_{\text{corr}}^x\rangle}{dt} = \frac{\rho BV_y}{V_x}. \quad (90)$$

We may then in turn compute the interaction entropy rate, again in the steady state, which due to a slightly cumbersome form we report in Appendix B. Together, these objects obey the bound in Eq. (87). These are illustrated in Fig. 4, where they are contrasted with the usual bipartite relation and, again for completeness, the bound with $\Delta\mathcal{S}_{\text{tot}}^x$ replaced with $\Delta\mathcal{S}_{\text{tot}}^x - \Delta\mathcal{S}_x^{\text{corr}}$. Importantly, we see that the previous conception of the generalized second law in such a system fails for some value of correlation ρ, holding for $\rho = 0$ where the system reduces to the usual bipartite model.

IX. DISCUSSION AND CONCLUSIONS

In this paper we have considered computations described in the framework of information dynamics. Using such a framework we have defined appropriate time-reversed computations and shown that the difference between each results in a measure of irreversibility that can be divided into computationally meaningful quantities: those that relate to the storage and transfer of information. By applying such a framework to physical systems we have shown that such a measure of computational irreversibility maps onto the total entropy production in stochastic thermodynamics revealing that not only can information processing lead to bounds on thermodynamics, as previously conceptualized, but that thermodynamics may be entirely described in terms of information processing.

Moreover, the framework based on information dynamics that we have presented is not restricted to the bipartite and multipartite systems previously considered in stochastic

thermodynamics. By dividing irreversibility along the lines of computational primitives, components of the total physical heat can be associated with irreversibility even when the physical heats exported by the subsystems are ambiguous. Further, we have shown that these irreversibilities based on storage and transfer are critically implicated in the resulting bounds that generalize the second law both when a bipartite assumption holds and when it does not. In both cases the bounds are achieved by correctly accounting for an interaction entropy production, which is always involved in its formulation and derived from a principle of conserved predictive capacity. Further, through the use of continuous dynamics with feedback and correlated noise we have presented a system where the physical heat exported by a subsystem is identifiable despite being governed by nonbipartite dynamics. Importantly, for this system, we have demonstrated that previous generalizations of the second law which account for feedback fail, while the bound arising from this formalism holds.

We hope that such a framework will allow discussion of irreversibility in systems where identification of individual heat flows are problematic and, moreover, allow discussion of thermodynamic processes in terms of distributed computation.

APPENDIX A: GENERALISED COMPUTATIONS

While not strictly necessary for the development of irreversibility measures, the intrinsic (forward) and time-reversed computations can be unified through the notion of a generalized intrinsic computation. To define such an object we appreciate that the basic structure of the intrinsic computation is the reduction in uncertainty of the next step in a time series, relative to an ensemble prior using the past sequence of events and a transition probability. And it is this notion that we seem to generalize based on the recognition of its structure in terms of uncertainties. First, we allow the prediction, or reduction of uncertainty, of some arbitrary event ε, taking values in some space $\mathcal{E}$, rather than merely the next step in the time series. This event must be computable with a sequence of states of variable X, that we write $\overrightarrow{X}$, taking values $\overrightarrow{x} \in \overrightarrow{\mathcal{X}}$. Consequently, we write $\varepsilon = f(\overrightarrow{x})$ such that $f : \overrightarrow{\mathcal{X}} \to \mathcal{E}$. We then wish to discuss the realizations of X that lead to the event ε such that we consider the preimage of ε,

$$f^{-1}(\varepsilon) = \{\overrightarrow{x} \in \overrightarrow{\mathcal{X}} | f(\overrightarrow{x}) = \varepsilon\}. \tag{A1}$$

Associated with this set of realizations $\overrightarrow{x}$ we have a prior distribution on $\overrightarrow{X}$, denoted $\mathbb{Q}(\overrightarrow{x} \in \mathcal{A})$, which represents our prior uncertainty in sets of $\overrightarrow{x}$. As such, the prior uncertainty in the specific event ε corresponds to $\mathcal{A} = f^{-1}(\varepsilon)$, i.e., $\mathbb{Q}[\overrightarrow{x} \in f^{-1}(\varepsilon)]$.

Finally, we consider an inference scheme that provides a posterior, which quantifies uncertainty in ε in the presence of some evidence $\overrightarrow{Z}$ taking values $\overrightarrow{z} \in \overrightarrow{\mathcal{Z}}$, where, like $\overrightarrow{X}$, $\overrightarrow{Z}$ consists of some time series. This inference scheme is represented with a probability measure $\mathbb{P}(\overrightarrow{x} \in \mathcal{A} | \overrightarrow{z})$, where $\mathcal{A} \subseteq \overrightarrow{\mathcal{X}}$, where, again, the set corresponding to the event ε is $\mathcal{A} = f^{-1}(\varepsilon)$, such that the uncertainty under inference scheme $\mathbb{P}$ and evidence $\overrightarrow{z}$ is $\mathbb{P}(\overrightarrow{x} \in f^{-1}(\varepsilon) | \overrightarrow{z})$.

Specifying all such quantities defines a tuple $\mathcal{C} = \{\overrightarrow{X}, f, \overrightarrow{Z}, \mathbb{P}, \mathbb{Q}\}$ that entirely characterizes the local computational signature of this generalized computation through the consideration of

$$c_{\mathcal{C}} = \ln \frac{d\mathbb{P}[\overrightarrow{x} \in f^{-1}(\varepsilon) | \overrightarrow{z}]}{d\mathbb{Q}[\overrightarrow{x} \in f^{-1}(\varepsilon)]}. \tag{A2}$$

In these terms we can identify the (forward) *intrinsic computation*, at time i, as being characterized by the tuple,

$$\mathcal{C}_{\{i,i+1\}} = \{X_{i+1:n}, X_{i+1}, \{X_{0:i}, Y_{0:i}\},$$
$$p_{X_{i+1:n}|X_{0:i},Y_{0:i}}(\cdot|\cdot), p_{X_{i+1:n}}(\cdot)\}. \tag{A3}$$

That is, the sequence of X from which the event is determined is the future states of the system, $X_{i+1:n}$. The predicted event is the trivial choice of the next state the system evolves to, $\varepsilon = f(x_{i+1:n}) = x_{i+1}$. The preimage $f^{-1}(\epsilon)$ is therefore the set of all paths $X_{i+1:n}$ that has $x_{i+1} = \varepsilon$, $f^{-1}(\varepsilon) = \{x_{i+1:n} \in \mathcal{X}_{i+1:n} : x_{i+1} = \varepsilon\}$. The evidence is the entirety of previous states of all available time series $\overrightarrow{Z} = \{X_{0:i}, Y_{0:i}\}$. Using these objects the prior belief is then that of the inherent uncertainty in the future paths of the system $\mathbb{Q}(x_{i+1:n}) = p(X_{i+1:n} = x_{i+1:n}) \equiv p_{X_{i+1:n}}(x_{i+1:n})$ and the inference strategy is provided by the same ensemble dynamics $\mathbb{P}(x_{i+1:n}|x_{0:i}, y_{0:i}) = p(X_{i+1:n} = x_{i+1:n}|X_{0:i} = x_{0:i}, Y_{0:i} = y_{0:i}) \equiv p_{X_{i+1:n}|X_{0:i},Y_{0:i}}(x_{i+1:n}|x_{0:i}, y_{0:i})$. One then returns the predictive capacity of the intrinsic computation by recognizing that

$$p_{X_{i+1:n}|X_{0:i},Y_{0:i}}(x_{i+1:n} \in f^{-1}(x_{i+1})|x_{0:i}, y_{0:i})$$
$$= p_{X_{i+1}|X_{0:i},Y_{0:i}}(x_{i+1}|x_{0:i}, y_{0:i})$$
$$= p(X_{i+1} = x_{i+1}|X_{0:i} = x_{0:i}, Y_{0:i} = y_{0:i}),$$
$$p_{X_{i+1:n}}(x_{i+1:n} \in f^{-1}(x_{i+1}))$$
$$= p_{X_{i+1}}(x_{i+1}) = p(X_{i+1} = x_{i+1}). \tag{A4}$$

In these terms one can define other information-theoretic quantities with other choices. For example, the predictive information, closely related to the excess entropy, results from the following choice, in the limit $n - i \to \infty, i \to \infty$,

$$\mathcal{C}_{\{i,i+1\}}^{\text{pred}} = \{X_{i+1:n}, X_{i+1:n}, X_{0:i}, p_{X_{i+1:n}|X_{0:i}}(\cdot|\cdot), p_{X_{i+1:n}}(\cdot)\}. \tag{A5}$$

In contrast, the definition of the time-reversed computation emerges through the tuple,

$$\mathcal{C}_{\{i,i+1\}}^{\dagger} = \Big\{X_{i:0}, X_i, \{X_{n:i+1}, Y_{n:i+1}\},$$
$$p_{X_{n-i:n}|X_{0:n-i-1},Y_{0:n-i-1}}^{\dagger}(\cdot, \cdot), p_{X_{i:0}}(\cdot)\Big\}. \tag{A6}$$

That is, the sequence from which the predicted event is computed is the time-reversed sequence of past states $\overrightarrow{X} = X_{i:0}$, the event is simply the "next" state in this sequence, $f(X_{i:0}) = X_i$ and the evidence sequence is the collection of time-reversed future states $\overrightarrow{Z} = \{X_{n:i+1}, Y_{n:i+1}\}$. Using these objects the prior distribution $\mathbb{Q}_{\overrightarrow{X}}$ is, as before, the ensemble probability of the sequence $\overrightarrow{X}$, i.e., $p_{X_{n-i:n}^{\dagger}}(x_{i:0}) = p_{X_{i:0}}(x_{i:0})$.

Finally, the inference strategy utilizes the time-reversed dynamics $p^{\dagger}_{X_{n-i:n}|X_{0:n-i-1},Y_{0:n-i-1}}(\cdot,\cdot)$.

APPENDIX B: TREATMENT OF THE CORRELATED COUPLED ORNSTEIN-UHLENBECK PROCESSES

The Fokker-Planck equation in Eq. (78) has a stationary solution of the form

$$p^s(x,y) = (2\pi|\Psi|)^{-\frac{1}{2}}\exp[\mathbf{z}^T\Psi^{-1}\mathbf{z}], \qquad (B1)$$

where

$$\mathbf{z} = \begin{pmatrix}\mathbf{x}\\\mathbf{y}\end{pmatrix}; \quad \Psi = \begin{pmatrix}\Psi_{xx} & \Psi_{xy}\\\Psi_{yx} & \Psi_{yy}\end{pmatrix}, \qquad (B2)$$

and

$$\Psi_{xx} = \frac{(BC-AD)V_x^2 + 2BD\rho V_x V_y - D^2 V_x^2 - B^2 V_y^2}{2(A+D)(AD-BC)},$$

$$\Psi_{yy} = \frac{(BC-AD)V_y^2 + 2CA\rho V_x V_y - A^2 V_y^2 - C^2 V_x^2}{2(A+D)(AD-BC)},$$

$$(B3)$$

$$\Psi_{xy} = \Psi_{yx} = \frac{CDV_x^2 - 2AD\rho V_x V_y + ABV_y^2}{2(A+D)(AD-BC)}.$$

Utilizing such a solution with Eq. (68) by first converting to Itō form, replacing $\langle dx\rangle$ with $Axdt + Bydt$ then averaging over $p^s(x,y)$ yields the interaction entropy production

$$\frac{d\langle\Delta\mathcal{S}^{xy}_{\text{int}}\rangle}{dt} = \frac{\rho\{CV_x^2 - V_y[(A-D)\rho V_x + BV_y]\}\{\rho CV_x^2 - V_y[(A-D)V_x + \rho BV_y]\}}{(A+D)(\rho^2-1)V_x^2 V_y^2}. \qquad (B4)$$

APPENDIX C: CALCULATION OF TRANSFER ENTROPY RATES OF THE CORRELATED COUPLED ORNSTEIN-UHLENBECK PROCESSES

In this Appendix we consider the coupled stochastic differential equations,

$$dx = Axdt + Bydt + V_x dW_x,$$
$$dy = Cxdt + Dydt + V_y dW_y, \qquad (C1)$$

with noise properties $\langle dW_x(t)dW_x(t')\rangle = \langle dW_y(t)dW_y(t')\rangle = \delta(t-t')dt$ and $\langle dW_x(t)dW_y(t')\rangle = \rho\delta(t-t')dt$ with $\rho \in [-1,1]$. This can be equivalently constructed by the following SDEs:

$$dx = Axdt + Bydt + \sqrt{1-\rho^2}V_x dW_1 + \rho V_x dW_2,$$
$$dy = Cxdt + Dydt + V_y dW_2, \qquad (C2)$$

with properties $dW_1(t)dW_2(t') = 0$. This leads to a diffusion matrix

$$\begin{bmatrix} \frac{V_x^2}{2} & \frac{\rho V_x V_y}{2} \\ \frac{\rho V_x V_y}{2} & \frac{V_y^2}{2} \end{bmatrix}. \qquad (C3)$$

This is a Gaussian process, since the noise is Gaussian and the driving terms are linear. Importantly, the coarse-grained process in X is also Gaussian. Critically, for stationary Gaussian processes the entropy rate may be written [16]

$$\dot{h}_x = \lim_{dt\to 0}\frac{1}{2dt}\ln(2\pi e) + \frac{1}{4\pi}\int_{-\pi/dt}^{\pi/dt}\ln\Sigma(\omega)d\omega, \qquad (C4)$$

such that, under certain circumstances, the transfer entropy is given by

$$T_{y\to x} = \dot{h}_x - \dot{h}_{x|\{y\}} = -\frac{1}{4\pi}\int_{-\infty}^{\infty}\ln\frac{\Sigma_{x|\{y\}}(\omega)}{\Sigma_x(\omega)}d\omega, \qquad (C5)$$

where $\Sigma_x(\omega)$ is the Fourier transform of the auto correlation function, $R_{xx}(\tau) = \mathbb{E}_x[x(t)x^*(t+\tau)]$, and $\Sigma_{x|\{y\}}(\omega)$ is the Fourier transform of the correlation function under the measure with knowledge of Y, i.e., $p[x_{t_0}^t|x_{t_0},\{y_{t_0}^t\}]$, $R_{xx|\{y\}}(\tau) =$ $\mathbb{E}_{x|\{y\}}[x(t)x^*(t+\tau)]$. Importantly, $\Sigma_{x|\{y\}}(\omega)$ can only be computed for the $\rho = 0$ case, differing from the general $\rho \neq 0$ case that we are considering.

These quantities, $\Sigma_{x|\{y\}}(\omega)$ and $\Sigma_x(\omega)$ (with $\rho = 0$ in the former) can be derived by Fourier transforming the SDEs. We write $\dot{x} = dx/dt$ and $\eta_x = V_x dW_x/dt$, such that we reformulate as Langevin equations, so that

$$i\omega\hat{x} = A\hat{x} + B\hat{y} + \hat{\eta}_x,$$
$$i\omega\hat{y} = C\hat{x} + D\hat{y} + \hat{\eta}_y. \qquad (C6)$$

Since we can only identify $\Sigma_{x|\{y\}}(\omega)$ for $\rho = 0$, we separate the computation of the transfer entropy rate into two components:

$$\dot{T}_{y\to x} = \dot{h}_x - \dot{h}_{x|\{y\}} = \underbrace{\dot{h}_x - \dot{h}^{\rho=0}_{x|\{y\}}}_{1} + \underbrace{\dot{h}^{\rho=0}_{x|\{y\}} - \dot{h}_{x|\{y\}}}_{2}. \qquad (C7)$$

First, we consider calculation one and so calculate $\Sigma_x(\omega)$ in the general case, $\rho \neq 0$, and $\Sigma_{x|\{y\}}(\omega) = \Sigma^{\rho=0}_{x|\{y\}}(\omega)$ in the $\rho = 0$ case.

To consider $\Sigma^{\rho=0}_{x|\{y\}}(\omega)$ we wish to consider the autocorrelation $\langle x'(t)x'(t+\tau)\rangle$, where $x' = x - \langle x|y\rangle$. To do so we follow Ref. [16] and consider not only the evolution of $\dot{x}$ and thus $\hat{x}$ but also $\langle\dot{x}|y\rangle$, which follows

$$d\langle x|y\rangle = A\langle x|y\rangle dt + Bydt,$$
$$i\omega\langle\hat{x}|y\rangle = A\langle\hat{x}|y\rangle + B\hat{y}. \qquad (C8)$$

Now, by the Wiener-Khinchin theorem,

$$\Sigma_{xy}(\omega) = \mathcal{F}\{R_{xy}(\tau)\} = \mathcal{F}\{\mathbb{E}[x(t)y^*(t+\tau)]\} = \mathbb{E}[\hat{x}\hat{y}^*], \qquad (C9)$$

where $\mathcal{F}$ is the Fourier transform.

Consequently,

$$\Sigma^{\rho=0}_{x|\{y\}}(\omega) = \mathcal{F}[\langle x'(t)x'(t+\tau)\rangle] = \langle|\hat{x} - \langle\hat{x}|y\rangle|^2\rangle$$

$$= \frac{\mathbb{E}[\hat{\eta}_x\hat{\eta}_x^*]}{(i\omega-A)(-i\omega-A)} = \frac{V_x^2}{\omega^2+A^2}. \qquad (C10)$$

Solving for $\hat{x}$, independently of y, allows computation of $\Sigma_x(\omega)$, where

$$\hat{x} = \frac{(i\omega - A)(i\omega - D)}{(i\omega - A)(i\omega - D) - BC}[B\hat{\eta}_y + (i\omega - D)\hat{\eta}_x], \tag{C11}$$

such that

$$\Sigma_x = \mathbb{E}[\hat{x}\hat{x}^*] = \left|\frac{(i\omega - A)(i\omega - D)}{(i\omega - A)(i\omega - D) - BC}\right|^2 \frac{1}{(A^2 + \omega^2)(D^2 + \omega^2)}$$

$$\times [(D^2 + \omega^2)\mathbb{E}[\hat{\eta}_x\hat{\eta}_x^*] + B^2\mathbb{E}[\hat{\eta}_y\hat{\eta}_y^*] - 2BD(\mathbb{E}[\hat{\eta}_x\hat{\eta}_y^*] + \mathbb{E}[\hat{\eta}_x^*\hat{\eta}_y]) + iB\omega(\mathbb{E}[\hat{\eta}_x\hat{\eta}_y^*] - \mathbb{E}[\hat{\eta}_x^*\hat{\eta}_y])]. \tag{C12}$$

Now, similarly,

$$\mathbb{E}[\hat{\eta}_x\hat{\eta}_y^*] = \mathcal{F}\{c(\eta_x(t), \eta_y^*(t + \tau))\} = \frac{V_x V_y dW_x(t)dW_y(t)}{dt}\int_{-\infty}^{\infty}\delta(\tau)e^{-i\omega\tau}d\tau = \frac{V_x V_y dW_x(t)dW_y(t)}{dt}. \tag{C13}$$

That is, V_x^2 for auto-correlation terms and $\rho V_x V_y$ for cross correlation terms. Consequently, we have

$$\mathbb{E}[\hat{\eta}_x\hat{\eta}_x^*] = V_x^2, \quad \mathbb{E}[\hat{\eta}_y\hat{\eta}_y^*] = V_y^2, \quad \mathbb{E}[\hat{\eta}_x\hat{\eta}_y^*] = \mathbb{E}[\hat{\eta}_x^*\hat{\eta}_y] = \rho V_x V_y, \tag{C14}$$

and so

$$\Sigma_x = \left|\frac{(i\omega - A)(i\omega - D)}{(i\omega - A)(i\omega - D) - BC}\right|^2 \frac{1}{(A^2 + \omega^2)(D^2 + \omega^2)}[(D^2 + \omega^2)V_x^2 + B^2 V_y^2 - 2\rho BDV_x V_y]. \tag{C15}$$

Looking to the integral in Eq. (C4) we have

$$-\frac{1}{4\pi}\int_{-\infty}^{\infty}\ln\frac{\Sigma_{x|\{y\}}^{\rho=0}(\omega)}{\Sigma_x(\omega)}d\omega = \frac{1}{4\pi}\int_{-\infty}^{\infty}\ln\left|\frac{(i\omega - A)(i\omega - D)}{(i\omega - A)(i\omega - D) - BC}\right|^2 - \frac{1}{4\pi}\int_{-\infty}^{\infty}\ln\frac{V_x^2(D^2 + \omega^2)}{(D^2 + \omega^2)V_x^2 + B^2 V_y^2 - 2\rho BDV_x V_y}. \tag{C16}$$

The first of these is a known vanishing integral [16]. For the second we exploit the integral

$$\int_0^{\infty}\ln\frac{x^2 + a^2}{x^2 + b^2}dx = \pi(a - b) \tag{C17}$$

to arrive at

$$\dot{h}_x - \dot{h}_{x|\{y\}}^{\rho=0} = \frac{|D|}{2}\left[\sqrt{1 + \frac{BV_y}{DV_x}\left(\frac{BV_y}{DV_x} - 2\rho\right)} - 1\right], \tag{C18}$$

but this is only half of the solution. Next we need to consider $\dot{h}_{x|\{y\}}^{\rho=0} - \dot{h}_{x|\{y\}}$. To do this we consider a different approach.

Since the system is linear/Gaussian, the entropy rate is entirely encoded by the variance in the (dt) regime. That is, we have

$$\dot{h}dt = \tfrac{1}{2}\ln(2\pi e\sigma^2 dt) + O(dt^2). \tag{C19}$$

To calculate this we explicitly describe an expression for the solution to the original Langevin equations (which, incidentally, is of closed form when $C = 0$). We do this individually for both X and Y by writing

$$\frac{d}{dt}[e^{-At}x(t)] = [By(t) + \eta_x(t)]e^{-At}, \quad \frac{d}{dt}[e^{-Dt}y(t)] = [Cx(t) + \eta_y(t)]e^{-Dt}, \tag{C20}$$

such that

$$x(t) = x(0)e^{At} + \int_0^t e^{A(t-t')}By(t')dt' + \int_0^t e^{A(t-t')}\eta_x(t')dt',$$

$$y(t) = y(0)e^{Dt} + \int_0^t e^{D(t-t')}Cx(t')dt' + \int_0^t e^{D(t-t')}\eta_y(t')dt'. \tag{C21}$$

We then substitute $y(t)$ in for $y(t')$ such that

$$x(t) = x(0)e^{At} + \int_0^t e^{A(t-t')}B\left[y(0)e^{Dt'} + \int_0^{t'} e^{D(t'-t'')}Cx(t'')dt'' + \int_0^{t'} e^{D(t'-t'')}\eta_y(t'')dt''\right]dt' + \int_0^t e^{A(t-t')}\eta_x(t')dt'. \tag{C22}$$

When $C \neq 0$ it is not of closed form, but adequate for our purposes. From this we want to find $\langle(x(t) - \langle x(t)\rangle)^2\rangle$, where the average $\langle x\rangle$ is shorthand for $\langle x(t)|x(0), y(0)\rangle$. We achieve this by noting

$$\left\langle\int_0^t f(t')\eta(t')dt'\right\rangle = \int_0^t f(t')\langle\eta(t')\rangle dt' = 0, \tag{C23}$$

since we have $\langle \eta(t') \rangle = 0$. As such we have

$$x(t) - \langle x(t) \rangle = B \int_0^t e^{A(t-t')} dt' \int_0^{t'} dt'' e^{D(t'-t'')} \eta_y(t'') + \int_0^t dt' e^{A(t-t')} \eta_x(t'), \tag{C24}$$

safely independent of C. Consequently, the variance is then given by

$$\langle (x(t) - \langle x(t) \rangle)^2 \rangle = \left\langle \int_0^t dt' \int_0^t dt'' e^{A(2t-t'-t'')} \eta_x(t') \eta_x(t'') \right\rangle + 2B \left\langle \int_0^t dt' \int_0^{t'} dt'' \int_0^t dt''' e^{A(2t-t'-t''')+D(t'-t'')} \eta_x(t''') \eta_y(t'') \right\rangle$$

$$+ B^2 \left\langle \int_0^t dt' \int_0^{t'} dt'' \int_0^t dt''' \int_0^{t'''} dt'''' e^{A(2t-t'-t''')+D(t'+t'''-t''-t'''')} \eta_y(t'') \eta_y(t'''') \right\rangle$$

$$= \int_0^t dt' \int_0^t dt'' e^{A(2t-t'-t'')} \langle \eta_x(t') \eta_x(t'') \rangle + 2B \int_0^t dt' \int_0^{t'} dt'' \int_0^t dt''' e^{A(2t-t'-t''')+D(t'-t'')}$$

$$\times \langle \eta_x(t''') \eta_y(t'') \rangle + B^2 \int_0^t dt' \int_0^{t'} dt'' \int_0^t dt''' \int_0^{t'''} dt'''' e^{A(2t-t'-t''')+D(t'+t'''-t''-t'''')} \langle \eta_y(t'') \eta_y(t'''') \rangle. \tag{C25}$$

Now we identify

$$\langle \eta_x(t') \eta_x(t'') \rangle = \frac{V_x^2}{2} \delta(t'' - t'), \quad \langle \eta_y(t') \eta_y(t'') \rangle = \frac{V_y^2}{2} \delta(t'' - t'), \quad \langle \eta_x(t') \eta_y(t'') \rangle = \frac{\rho V_x V_y}{2} \delta(t'' - t'), \tag{C26}$$

and also the usual sifting property

$$\int_0^t \left[\int_0^{t'} f(t''', t'') \delta(t''' - t'') dt''' \right] dt'' = \int_0^t 1_{[0,t']}(t'') f(t'', t') dt'' = \int_0^{\min(t,t')} f(t'', t') dt'' = \int_0^{t'} f(t'', t') dt'' \quad (t > t'). \tag{C27}$$

Consequently,

$$\langle (x(t) - \langle x(t) \rangle)^2 \rangle = \frac{V_x^2}{2} \int_0^t dt' e^{2A(t-t')} + \rho B V_x V_y \int_0^t dt' \int_0^{t'} dt''' e^{A(2t-t't''')+D(t'-t''')}$$

$$+ B^2 \frac{V_y^2}{2} \int_0^t dt' \int_0^t dt''' \int_0^{\min(t',t''')} dt'' e^{A(2t-t'-t''')+D(t'+t'''-2t'')}$$

$$= \frac{V_x^2}{2} \frac{e^{2At} - 1}{2A} + \rho B V_x V_y \frac{A - D + (A+D)e^{2At} - 2Ae^{(A+D)t}}{2(A^3 - AD^2)}$$

$$+ B^2 \frac{V_y^2}{2} \left[\int_0^t dt' \int_{t'}^t dt''' \int_0^{t'} dt'' + \int_0^t dt''' \int_{t'''}^t dt' \int_0^{t'''} dt'' \right] e^{A(2t-t'-t''')+D(t'+t'''-2t'')}$$

$$= \frac{V_x^2}{2} \frac{e^{2At} - 1}{2A} + \rho B V_x V_y \frac{A - D + (A+D)e^{2At} - 2Ae^{(A+D)t}}{2(A^3 - AD^2)}$$

$$+ B^2 \frac{V_y^2}{2} \frac{D^2(e^{2At} - 1) + A^2(e^{2Dt} - 1) + AD(2 + e^{2At} + e^{2Dt} - 4e^{(A+D)t})}{2AD(A+D)(A-D)^2}. \tag{C28}$$

The first term is $O(t)$, the second $O(t^2)$ and the third $O(t^3)$. Examining the form of Eq. (C19) we need second order contributions in σ^2 to capture first-order terms in $\dot{h} dt$ and so we identify, expanding the exponentials,

$$\langle (x(t) - \langle x(t) \rangle)^2 \rangle = \frac{V_x^2 t}{2} \left(1 + At + \rho B \frac{V_y}{V_x} t \right) + O(t^3), \tag{C29}$$

such that

$$\dot{h}_{x|\{y\}} dt = \frac{1}{2} \ln \left[\pi e V_x^2 dt \right] + \frac{A}{2} dt + \frac{\rho B V_y}{2 V_x} dt, \tag{C30}$$

and therefore

$$\dot{h}_{x|\{y\}}^{\rho=0} - \dot{h}_{x|\{y\}} = -\frac{\rho B V_y}{2 V_x}. \tag{C31}$$

Putting this together with Eqs. (C7) and (C18) we finally get the result

$$\dot{T}_{y \to x} = \frac{|D|}{2} \left[\sqrt{1 + \frac{BV_y}{DV_x} \left(\frac{BV_y}{DV_x} - 2\rho \right)} - \left(1 + \rho \frac{BV_y}{|D|V_x} \right) \right], \tag{C32}$$

with the appropriate $\dot{T}_{x \to y}$ result found by simply exchanging variables $V_x \leftrightarrow V_y$, $A \leftrightarrow D$, $B \leftrightarrow C$, due to symmetry. We note the lack of dependence on A and C.

---

[1] D. J. Evans, E. G. D. Cohen, and G. P. Morriss, Phys. Rev. Lett. **71**, 2401 (1993).

[2] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[3] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).

[4] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[5] C. Jarzynski, Phys. Rev. E **56**, 5018 (1997).

[6] U. Seifert, Phys. Rev. Lett. **95**, 040602 (2005).

[7] U. Seifert, Eur. Phys. J. B **64**, 423 (2008).

[8] U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).

[9] A. C. Barato and U. Seifert, Phys. Rev. E **90**, 042150 (2014).

[10] D. Mandal and C. Jarzynski, Proc. Natl. Acad. Sci. USA **109**, 11641 (2012).

[11] S. Deffner and C. Jarzynski, Phys. Rev. X **3**, 041003 (2013).

[12] A. Chapman and A. Miyake, Phys. Rev. E **92**, 062125 (2015).

[13] T. McGrath, N. S. Jones, P. R. ten Wolde, and T. E. Ouldridge, Phys. Rev. Lett. **118**, 028101 (2017).

[14] J. M. Horowitz and S. Vaikuntanathan, Phys. Rev. E **82**, 061120 (2010).

[15] J. M. Horowitz and M. Esposito, Phys. Rev. X **4**, 031015 (2014).

[16] J. M. Horowitz and H. Sandberg, New J. Phys. **16**, 125007 (2014).

[17] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, Nat. Phys. **11**, 131 (2015).

[18] M. Esposito and G. Schaller, Europhys. Lett. **99**, 30003 (2012).

[19] T. Sagawa and M. Ueda, Phys. Rev. Lett. **100**, 080403 (2008).

[20] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010).

[21] T. Sagawa and M. Ueda, Phys. Rev. E **85**, 021104 (2012).

[22] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, Nat. Phys. **6**, 988 (2010).

[23] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Phys. Rev. Lett. **109**, 120604 (2012).

[24] P. Strasberg, G. Schaller, T. Brandes, and M. Esposito, Phys. Rev. Lett. **110**, 040601 (2013).

[25] D. Abreu and U. Seifert, Phys. Rev. Lett. **108**, 030601 (2012).

[26] T. Sagawa and M. Ueda, Phys. Rev. Lett. **102**, 250602 (2009).

[27] C. P. Kempes, D. Wolpert, Z. Cohen, and J. Pérez-Mercader, Philos. Trans. R. Soc. London A **375**, 20160343 (2017).

[28] D. Wolpert, *APS Meeting Abstracts* (APS 2016), p. A40.003.

[29] M. Esposito and C. V. den Broeck, Europhys. Lett. **95**, 40004 (2011).

[30] M. Prokopenko and I. Einav, Phys. Rev. E **91**, 062143 (2015).

[31] M. Prokopenko and J. T. Lizier, Sci. Rep. **4**, 5394 (2014).

[32] J. P. Crutchfield, Physica D **75**, 11 (1994).

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

[34] C. Langton, in *Emergent Computation*, edited by S. Forest (MIT, 1991).

[35] C. G. Langton, Physica D: Nonlin. Phenom. **42**, 12 (1990).

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

[37] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, in *Guided Self-Organization: Inception*, Emergence, Complexity and Computation, Vol. 9, edited by M. Prokopenko (Springer, Berlin/ Heidelberg, 2014), pp. 115–158.

[38] 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).

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

[40] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko, J. Comput. Neurosci. **30**, 85 (2010).

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

[42] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Info. Sci. **208**, 39 (2012).

[43] 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/ Heidelberg, 2010), pp. 38–49.

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

[45] 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\ Heidelberg, 2014), pp. 61–86.

[46] S. Dasgupta, F. Wörgötter, and P. Manoonpong, Evol. Syst. **4**, 235 (2013).

[47] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Theory Biosci. **131**, 193 (2012).

[48] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, J. Theor. Biol. **218**, 1 (2002).

[49] I. Couzin, R. James, D. Croft, and J. Krause, in *Fish Cognition and Behavior*, Fish and Aquatic Resources, edited by C. Brown, K. Laland, and J. Krause (Blackwell Publishing, London, 2006), pp. 166–185.

[50] S. Bikhchandani, D. Hirshleifer, and I. Welch, J. Pol. Eco. **100**, 992 (1992).

[51] B. G. Galef, Jr. and L.-A. Giraldeau, Anim. Behav. **61**, 3 (2001).

[52] D. Sumpter, J. Buhl, D. Biro, and I. Couzin, Theory Biosci. **127**, 177 (2008).

[53] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems* (Oxford University Press, Oxford, 1999).

[54] I. D. Couzin, Trends Cognit. Sci. **13**, 36 (2009).

[55] L. Albantakis, A. Hintze, C. Koch, C. Adami, and G. Tononi, PLoS Comput. Biol. **10**, e1003966 (2014).

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

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

[58] S. Butail, F. Ladu, D. Spinello, and M. Porfiri, Entropy **16**, 1315 (2014).

[59] N. Orange and N. Abaid, Eur. Phys. J.: Spec. Top. **224**, 3279 (2015).

[60] E. Crosato, L. Jiang, V. Lecheval, J. T. Lizier, X. R. Wang, P. Tichit, G. Theraulaz, and M. Prokopenko, Swarm Intell. (2018).

[61] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993).

[62] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, in *Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI), Winchester, UK*, edited by S. Bullock, J. Noble, R. Watson, and M. A. Bedau (MIT Press, Cambridge, MA, 2008), pp. 374–381.

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

[64] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich, Eur. Phys. J. B **63**, 329 (2008).

[65] R. Der and G. Martius, *The Playful Machine: Theoretical Foundation and Practical Realization of Self-Organizing Robots*, Cognitive Systems Monographs (Springer, Berlin, 2012).

[66] G. Martius, J. Herrmann, and R. Der, in *Advances in Artificial Life* (2007), pp. 766–775.

[67] G. Martius and J. M. Herrmann, in *Proceedings of the 11th International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, SAB'10 (Springer-Verlag, Paris, France, 2010), pp. 50–61.

[68] N. Ay, H. Bernigau, R. Der, and M. Prokopenko, Theory Biosci. **131**, 161 (2012).

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

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

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

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

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

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

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

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

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

[78] Robert Fano, *Transmission of Information: A Statistical Theory of Communications* (MIT Press, Cambridge, MA, 1961).

[79] G. E. Crooks, J. Stat. Phys. **90**, 1481 (1998).

[80] 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, 2013), pp. 3–56.

[81] R. E. Spinney, J. T. Lizier, and M. Prokopenko, Phys. Rev. E **94**, 022135 (2016).

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

[83] B. H. Shargel, J. Phys. A: Math. Theor. **43**, 135002 (2010).

[84] M. Bauer and F. Cornu, J. Phys. A: Math. Theor. **48**, 015008 (2015).

[85] P. G. Bergmann and J. L. Lebowitz, Phys. Rev. **99**, 578 (1955).

[86] M. Colangeli, C. Maes, and B. Wynants, J. Phys. A: Math. Theor. **44**, 095001 (2011).

[87] B. Derrida, J. Stat. Mech. (2007) P07023.

[88] J. L. Lebowitz, Phys. Rev. **114**, 1192 (1959).

[89] J. L. Lebowitz and H. Spohn, J. Stat. Phys. **95**, 333 (1999).

[90] C. Maes and K. Netočný, J. Stat. Phys. **110**, 269 (2003).

[91] H. Tasaki, arXiv:0706.1032[cond-mat] (2007).

[92] M. Esposito and C. Van den Broeck, Phys. Rev. Lett. **104**, 090601 (2010).

[93] I. Prigogine, *Thermodynamic of Irreversible Processes* (Interscience Publishers, New York, 1967).

[94] R. Chetrite and K. GawÈdzki, Commun. Math. Phys. **282**, 469 (2008).

[95] M. Esposito and C. Van den Broeck, Phys. Rev. E **82**, 011143 (2010).

[96] R. E. Spinney and I. J. Ford, Phys. Rev. Lett. **108**, 170603 (2012).

[97] C. Van den Broeck and M. Esposito, Phys. Rev. E **82**, 011144 (2010).

[98] T. Hatano and S.-i. Sasa, Phys. Rev. Lett. **86**, 3463 (2001).

[99] T. Speck and U. Seifert, J. Phys. A: Math. Gen. **38**, L581 (2005).

[100] Y. Oono and M. Paniconi, Prog. Theor. Phys. Suppl. **130**, 29 (1998).

[101] S. Ito and T. Sagawa, Phys. Rev. Lett. **111**, 180603 (2013).

[102] S. Ito and T. Sagawa, Nat. Commun. **6**, 7498 (2015).

[103] D. Hartich, A. C. Barato, and U. Seifert, J. Stat. Mech. (2014) P02016.

[104] D. Hartich, A. C. Barato, and U. Seifert, Phys. Rev. E **93**, 022116 (2016).

[105] A. E. Allahverdyan, D. Janzing, and G. Mahler, J. Stat. Mech. (2009) P09011.

[106] Ken Sekimoto, *Stochastic Energetics*, Lecture Notes in Physics, Vol. 799 (Springer, Berlin, 2010).

[107] K. Sekimoto, Prog. Theor. Phys. Suppl. **130**, 17 (1998).

[108] K. Sekimoto, J. Phys. Soc. Jpn. **66**, 1234 (1997).

[109] R. E. Spinney and I. J. Ford, Phys. Rev. E **85**, 051113 (2012).

[110] R. E. Spinney, Ph.D. thesis, UCL, University College London, 2012.

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