Title
Fluctuations When Driving Between Nonequilibrium Steady States

Permalink
https://escholarship.org/uc/item/1491w9jp

Journal
Journal of Statistical Physics, 168(4)

ISSN
0022-4715

Authors
Riechers, Paul M
Crutchfield, James P

Publication Date
2017-08-01

DOI
10.1007/s10955-017-1822-y

Peer reviewed

Santa Fe Institute Working Paper 16-10-023
arxiv.org:1610.09444 [cond-mat.stat-mech]

Fluctuations When Driving Between Nonequilibrium Steady States

Paul M. Riechers* and James P. Crutchfield†
Complexity Sciences Center, Department of Physics
University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: May 13, 2017)

Maintained by environmental fluxes, biological systems are thermodynamic processes that operate far from equilibrium without detailed-balanced dynamics. Yet, they often exhibit well defined nonequilibrium steady states (NESSs). More importantly, critical thermodynamic functionality arises directly from transitions among their NESSs, driven by environmental switching. Here, we identify the constraints on excess heat and dissipated work necessary to control a system that is kept far from equilibrium by background, uncontrolled "housekeeping" forces. We do this by extending the Crooks fluctuation theorem to transitions among NESSs, without invoking an unphysical dual dynamics. This and corresponding integral fluctuation theorems determine how much work must be expended when controlling systems maintained far from equilibrium. This generalizes thermodynamic feedback control theory, showing that Maxwellian Demons can leverage mesoscopic-state information to take advantage of the excess energetics in NESS transitions. We also generalize an approach recently used to determine the work dissipated when driving between functionally relevant configurations of an active energy-consuming complex system. Altogether, these results highlight universal thermodynamic laws that apply to the accessible degrees of freedom within the effective dynamic at any emergent level of hierarchical organization. By way of illustration, we analyze a voltage-gated sodium ion channel whose molecular conformational dynamics play a critical functional role in propagating action potentials in mammalian neuronal membranes.

PACS numbers: 05.70.Ln 05.20.-y 05.45.-a 02.50.Ey
Keywords: stochastic thermodynamics, fluctuation theorem, nonequilibrium, neuronal ion channel

I. INTRODUCTION

The sun shines; ATP is abundant; power is supplied. These are the generous settings in which we find many complex biological systems, buoyed steadily out of equilibrium by energy fluxes in their environment. The resulting steady-state dynamics exhibit various types of directionality, including periodic oscillations and macroscopic thermodynamic functionality. These behaviors contrast rather sharply with the deathly isotropy of equilibrium detailed-balanced dynamics—where fluxes are absent and state transition rates depend only on relative asymptotic state-occupation probabilities.

Detailed balance and its implied dynamical reversibility, though, are common tenets of equilibrium thermodynamics and statistical mechanics. They are technically necessary when applying much of the associated theory relevant to equilibrium and reversible (e.g., quasistatic) transitions between equilibrium macrostates [1]. Detailed balance is even assumed by several modern theorems that influence our understanding of the structure of fluctuations and the limitations on work performed far from equilibrium [2, 3]. The natural world, though, is replete with systems that violate detailed balance, such as small biological molecules constantly driven out of equilibrium through interactions with their biochemical environment [4, 5].

Far from happenstance and disruption, the probability currents through the effective state-space of these nonequilibrium systems enable crucial thermodynamic functionality [6, 7]. Even rare fluctuations play an important functional role [8–10]. While constant environmental pressure can drive a system into a nonequilibrium steady state (NESS), complex biological systems are often driven farther—far from even any NESS. Moreover, such system–environment dynamics involve feedback between system and environment states. Although many believe these facilitate the necessary complex processes that sustain life, their very nature seems to preclude most, if not all, hope of a universal theoretical framework for quantitative predictions. To ameliorate the roadblock, we present a consistent thermodynamics that is not only descriptive, but constructive, tractable, and predictive, even when irreversible dynamics transition between NESSs.

Beyond laying out the structure of fluctuations during transitions between NESSs, this thermodynamics sets the stage to understand how one level of organization gives way to another; cf. [11, Sec. 5.2]. In particular, the results herein enable the iterative renormalization of the nonequilibrium housekeeping background, which shows how a hierarchy of steady-state dynamics can be maintained; see Ref. [12, Ch. 8] for preliminary work in this direction. Said simply, at each level of hierarchical orga-

* pmriechers@ucdavis.edu
† chaos@ucdavis.edu

nization, controllable degrees of freedom are subject to universal thermodynamic laws that tie their fluctuations and functionality to dissipation at lower levels. Here, the emphasis is to characterize the energy dissipation due to controllable forces at any level—to disentangle the work that must be exerted in contradistinction to the entropy produced by uncontrolled forces that play out in the self-sustaining background.

A. Results

Nonequilibrium thermodynamics progressed markedly over the last two decades on at least two fronts. First, by taking the 'dynamics' in 'thermodynamics' seriously, fluctuation theorems (FTs) transformed previous inequalities, such as the classical Second Law of Thermodynamics, into subsuming equalities that exactly express the distribution of thermodynamic variations. (These have been derived by many authors now in a wide range of physical settings; see, e.g., Refs. [13] and [14] for lucid reviews.) Second, steady-state thermodynamics (SST) showed that NESSs play a role in nonequilibrium analogous to that of equilibrium macrostates in equilibrium. In this view heat decomposes into the *housekeeping heat* Q_{hk} needed to sustain NESSs and the *excess heat* Q_{ex} dissipated in transitions between them [15–17]. Bolstering SST, recent efforts generalized the Clausius inequality (describing excess heat produced beyond the change in system entropy) to smoothly driven transitions between NESSs [18]. Taken together, these results established an integral fluctuation theorem for the excess work in NESS transitions and, consequently, a generalized Second Law for excess entropy produced beyond housekeeping during driven NESS transitions.

The following extends SST by introducing several new FTs, highlighting correspondences between nonequilibrium and equilibrium relations. It also brings to the fore a fresh perspective on control and feedback, addressing (i) the dissipated work a controller must exert to influence a system when the controller only has access to a subset of the forces that drive the system and (ii) intrinsic feedback and the thermodynamic relevance of informative auxiliary variables besides measurement.

First, we provide detailed (i.e., nonintegrated) fluctuation theorems, rather than integral fluctuation theorems for driven NESS transitions. (Integral FTs follow directly, in any case.) Moreover, since these detailed FTs avoid using a nonphysical dual dynamics, this constrains experimentally accessible distributions of excess work W_{ex} exerted when controlling nonequilibrium systems. As part of this theoretical development, we jointly bound housekeeping and excess work distributions induced by any control protocol.

As a simple example, for time-symmetric driving we show that the joint probability of excess work and housekeeping heat respect the strong constraint:

$$\frac{\Pr(W_{\text{ex}}, Q_{\text{hk}})}{\Pr(-W_{\text{ex}}, -Q_{\text{hk}})} = e^{\beta W_{\text{ex}}} e^{\beta Q_{\text{hk}}} ~, \qquad (1)$$

when starting from a steady-state distribution. In a biological setting, Q_{hk} addresses the energetic cost of homeostasis, while W_{ex} addresses the inefficiency of adaptive response to new stimuli. In a control setting, W_{ex} is the work exerted for control that will be irretrievably dissipated when the system is driven from one steady state to another.

Our jointly-constrained fluctuation theorem Eq. (1) unifies and generalizes several important previous fluctuation theorems. For example, if no control is exerted to drive the system between steady states, then $W_{\text{ex}} = 0$ and we recover an exact fluctuation theorem regarding entropy production in a NESS:

$$\frac{\Pr(Q_{\text{hk}})}{\Pr(-Q_{\text{hk}})} = e^{\beta Q_{\text{hk}}} ~,$$

which is related to the transient FT for the dissipation function of Evans and Searles [19] and explains the original FTs postulated for NESSs [20–22]. (This is also reminiscent of a FT of Esposito and Van den Broeck [23] but, again, our results differ in that we avoid artificial dual dynamics.) At the opposite extreme, when the transitions are nonequilibrium (e.g., nonquasistatic) excursions between *equilibrium* steady states, then $Q_{\text{hk}} = 0$ and Eq. (1) reduces to the Crooks FT [24] for the excess work performed (beyond the change in free energy) via time-symmetric driving on a detailed-balanced system:

$$\frac{\Pr(W_{\text{ex}})}{\Pr(-W_{\text{ex}})} = e^{\beta W_{\text{ex}}} ~.$$

However, when control is exerted to drive a system *between nonequilibrium steady states*, then our jointly-constrained fluctuation theorem describes a new physical law not contained in the previous FTs.

More generally, we derive the detailed FTs for entropy production even when temperature varies in space and time. They are expressed in terms of excess environmental entropy production Ω and irreversibility Ψ, even when the irreversibility is "housekeeping" not strictly associated with heat. Moreover, we address the modifications needed for the fluctuation theorems when the system starts and ends out of steady state, through the inclusion of nonsteady-state additions to free energy.

We quantify a system's net *path irreversibility* Ψ with the accumulated violation of detailed balance in the effective dynamic. In the isothermal setting, for example, the irreversibility is the housekeeping heat, maintaining the system in its nonequilibrium dynamic: $\Psi = \beta Q_{\mathrm{hk}}$. Importantly, we can determine the minimum housekeeping heat without appealing to the system's Hamiltonian [25].

After the detailed FTs, we derive and discuss new integral FTs that generalize the results of Hatano and Sasa [16], as well as those of recent feedback control thermodynamics [3, 26]. We also discuss intrinsic feedback and the thermodynamic relevance of informative auxiliary variables besides measurement.

Our final theoretical contribution places novel constraints on the work dissipated W_{diss} (even when starting and ending far from steady state) in driving an active system between two functionally relevant collections of configurations, $\mathbf{I}$ and $\mathbf{II}$, via the control protocol $\mathbf{x}$. We find that:

$$\begin{aligned} &\langle W_{\mathrm{diss}} \rangle_{\Pr(\overrightarrow{s} | \mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})} \\ &\quad \geq \beta^{-1} \ln \frac{\Pr(\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})}{\Pr(\mathbf{II} \xrightarrow{\mathbf{x}^{\mathrm{R}}} \mathbf{I} \,|\, \mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})} - \beta^{-1} \langle \Psi \rangle_{\Pr(\overrightarrow{s} | \mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})} ~, \end{aligned}$$

where the dissipated work is averaged over all possible state paths $\overrightarrow{s}$ that start in $\mathbf{I}$ and end in $\mathbf{II}$, and $\mathbf{x}^{\mathrm{R}}$ is the reversed control protocol. This result suggests that an intelligent controller can leverage autonomous dissipative mechanisms in the housekeeping background to reduce its own work expenditure. We briefly discuss how this result generalizes recent efforts on the thermodynamics of replication [27–29], suggesting that the generalization is key to estimating energetic costs for realistic control of complex systems.

Extending SST in these ways reveals universal constraints on excess thermodynamic quantities—effective energies accessible above the housekeeping background. Looking forward, this allows one to analyze nondetailed-balanced stochastic dynamics—and thus contributes an understanding of the role of hierarchy—in the thermodynamics of replication [27] and the thermodynamics of learning [11]. Moreover, this identifies how complex, possibly intelligent, thermodynamic systems leverage (designed or intrinsic) irreversibility in their own state-space to harness energy and other thermodynamic resources from structured environments.

B. Synopsis

Section II sets up our approach, introducing notation, discussing input-dependent system dynamics, and establishing fundamental relationships among nonequilibrium thermodynamic quantities. Section III A introduces excess heat and excess work in analogy to classical heat and work. Ultimately though, the related excess environmental entropy production Ω discussed in § III B generalizes these to the case of temperature inhomogeneity over spacetime. Section III C demonstrates that path-induced entropies are the fundamental objects of nonequilibrium thermodynamics. In steady state, unaveraged path-induced entropies reduce to the steady-state surprisal ϕ. Deviations from the asymptotic surprisal contribute to a nonsteady-state additional free energy. All of these quantities play a central role in the subsequent development.

Before delving into irreversibility, though, we first address what is meant by reversibility. Therefore, § IV A and § IV B discuss detailed balance, microscopic reversibility, and the close relationship between them. Section IV C then introduces path dependence and reverse-path dependence and explains how together they yield a system's irreversibility Ψ.

With this laid out, § V A and § V C derive the detailed FTs in terms of excess environmental entropy production Ω and irreversibility Ψ. One sees that in the isothermal setting $\Psi = \beta Q_{\mathrm{hk}}$ and the excess entropy production is directly related to the excess work. This allows § V E to explain how these results extend SST.

Sections V F and V G finish our investigation of NESS FTs by deriving several integral FTs. This, in effect, extends feedback control, as developed in Refs. [3] and [26], to SST. We note that such environmental feedback is intrinsic to natural systems.

Section VI considers the work dissipated in driving an active system between two functionally relevant sets of configurations. This generalizes the theoretical basis previously introduced for self-replication and suggests that clever control protocols can leverage intrinsic irreversibility to achieve a desired influence while expending less energy.

For concreteness, § VII analyzes a simple but biologically important prototype system: voltage-gated sodium ion channels. These are complex macromolecules that violate detailed balance in order to perform critical biological functioning far from equilibrium. Finally, appendices discuss non-Markovian dynamics and comment on the bounds provided by integral fluctuation theorems for auxiliary variables.

II. DRIVEN STOCHASTIC DYNAMICS

We consider a classical system—the *system under study*—with time-dependent driving via environmentally determined parameters; e.g., time-dependent temperature, voltage, and piston position. Hence, the environmental control input X_t at time t, taking on values $x_t \in \mathcal{X}$, will typically be a vector object. The system under study is assumed to have a countable set $\boldsymbol{\mathcal{S}}$ of states. The random variable $\mathcal{S}_t$ for the state at time t takes on values $s_t \in \boldsymbol{\mathcal{S}}$. We assume that the environment's control value (current *input*) x and the system's physical state (current *state*) s are sufficient to determine the system's net effective energy—the *nonequilibrium potential* $\phi(x, s)$. Even with constant environmental input, the system dynamic need not be detailed balance.

A. Stochastic mesoscopic dynamics and induced state-distributions

We assume the current environmental input x determines the instantaneous stochastic transition dynamic over the system's observable mesoscopic states. However, that input can itself depend arbitrarily on all previous input and state history. That is, we assume that the $\boldsymbol{\mathcal{S}}$-to-$\boldsymbol{\mathcal{S}}$ transitions are instantaneously—that is, conditionally—Markovian given the input. Over time, though, different inputs induce different Markov chains over system states.

Note that the Markov assumption is common, although often implicit, and we follow this here to isolate the novel implications of nondetailed-balanced dynamics. Nevertheless, the results generalize to infinite Markov order by modeling system states as the observable output of many-to-one mappings of latent states of an input-controllable hidden Markov chain. Appendix A details this generalization.

We do not restrict the environment's driving process, allowing arbitrary non-Markovity, feedback, and nonstationarity. Thus, the joint system–environment dynamic can be non-Markovian even if the instantaneous system dynamic is. Such a setup is quite general, and so the results to follow extend others known for SST. We also follow stochastic thermodynamics in the use of (arbitrarily small) discrete-time steps. Nevertheless, it is usually easy to take the continuous-time limit. As, in fact, we do in the ion-channel example at the end.

Hence, the Markovian dynamic is described by a (possibly infinite) set of input-conditioned transition matrices over the state set $\boldsymbol{\mathcal{S}}$: $\{\mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x)}\}_{x\in\mathcal{X}}$, where $\mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x)}_{i,j} = \Pr(\mathcal{S}_t = s^j | \mathcal{S}_{t-1} = s^i, X_t = x)$ is the probability that the system is in state s^j at time t given that the system was in state s^i at time $t-1$ and the instantaneous environmental input controlling the system was x. Physically, the stochastic nature of these transitions typically stems from both the inherent information-producing chaotic dynamics of the system as well as the unpredictable interactions with the marginalized degrees of freedom of the thermalizing environment.

The Perron–Frobenius theorem guarantees that there is a stationary distribution $\boldsymbol{\pi_x}$ over states associated with each fixed input x. These are the state distributions associated with the system's nonequilibrium steady states (NESSs). For simplicity, and unless otherwise stated, we assume that a fixed input x eventually induces a unique NESS.

We denote probability distributions over the system states as bold Greek symbols; such as $\boldsymbol{\mu}$. We denote the state random variable $\mathcal{S}$ being distributed according to $\boldsymbol{\mu}$ via: $\mathcal{S} \sim \boldsymbol{\mu}$. It will often be convenient to cast $\boldsymbol{\mu}$ as a row-vector, in which case it appears as the bra $\langle \boldsymbol{\mu}|$. Putting this altogether, a sequence of driving inputs updates the state distribution as follows:

$$\begin{aligned}\langle \boldsymbol{\mu}_{t+n}| &= \langle \boldsymbol{\mu}_t| \, \mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x_{t:t+n})} \\ &= \langle \boldsymbol{\mu}_t| \, \mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x_t)} \mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x_{t+1})} \dots \mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x_{t+n-1})} ~.\end{aligned}$$

(Time indexing here and throughout is denoted by subscripts $t : t'$ that are left-inclusive and right-exclusive.) An infinite driving history $\overleftarrow{x} = \dots x_{-2}x_{-1}$ induces a distribution $\boldsymbol{\mu}(\overleftarrow{x})$ over the state space. The so-called *steady-state distribution* associated with the environmental drive value x, induced by tireless repetition of x, is:

$$\langle \boldsymbol{\pi_x}| = \lim_{n\to\infty} \langle \boldsymbol{\mu}_0| \left(\mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x)}\right)^n ~.$$

Usefully, $\boldsymbol{\pi_x}$ can also be found as the left eigenvector of $\mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x)}$ associated with the eigenvalue of unity [30]:

$$\langle \boldsymbol{\pi_x}| = \langle \boldsymbol{\pi_x}| \, \mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x)} ~. \tag{2}$$

The assumption that observable state-to-state transitions are instantaneously Markovian allows the state distribution $\boldsymbol{\mu}$ to summarize the causal relevance of the entire driving history $\overleftarrow{x}$.

It should be emphasized that this approach is quite general in addressing the dynamics and thermodynamics of realistic physical systems. While the approach is natural for many explicitly discrete-state models, it is also of fundamental importance for systems that are less obviously discrete. For example, many-bodied chaotic dynamical systems—typically modeled by nonlinear differential equations in continuous time and studied at the level of individual trajectories—are fully described by the linear Ruelle–Frobenius–Perron transition operator [31, 32] that evolves complex-valued *distributions* (in the sense of generalized functions) over the original state-space via some finite time-step. Our setup requires us to go beyond a single Ruelle–Frobenius–Perron operator, to utilize a full (possibly uncountable) *set* of Ruelle–Frobenius–Perron transition operators, each associated with a different control setting. The transition dynamic approach to nonequilibrium thermodynamics has the great advantage that the steady-state distribution $\boldsymbol{\pi_x}$ induced by any such fixed operator is simply an eigenstate of the operator, whereas this $\boldsymbol{\pi_x}$ often has a fractal support [21] in the phase-space representation typically employed in the classical analysis of differential equations. However, quantum mechanical considerations suggest that the physics of the system is fully described by the transitions among a countable basis—and, so, the delicate fractal nature of the phase-space representation of $\boldsymbol{\pi_x}$ is rather irrelevant to the thermodynamics, and perhaps indicative of a poor choice of representation for the present task. Finally, we remark that the transition operator approach to nonequilibrium thermodynamics lends itself to more easily reconciling quantum and classical results.

III. ENERGETICS AND ENTROPIES

For later comparison, we recount the basics of a statistical mechanics description of the thermodynamics of a system exchanging energy with a large environment, imposing fixed constraints indexed as x. The many-body Hamiltonian $\mathcal{H}(x)$ has energy eigenvalues $\{E(x,s)\}$, where s indexes the energy eigenstates. The canonical distribution is $\pi_x(s) = e^{-\beta[E(x,s)-F_{\mathrm{eq}}(x)]}$, at fixed x. This distribution is the equilibrium steady "state" associated with x, where $\beta^{-1} \equiv k_{\mathrm{B}}T$, T is the temperature of the macroscopic environment surrounding the system, and $F_{\mathrm{eq}}(x)$ is the associated equilibrium free energy.

A. Work, heat, and their excesses

Work W is environmentally driven energy change. Within one time-step it is given by [33]:

$$W[x_{n-1} \to x_n; s_{n-1}] = E(x_n, s_{n-1}) - E(x_{n-1}, s_{n-1}) \ .$$

Heat Q is the change in system energy due to its internal response to the environmental drive; e.g., a molecule's change in conformation. Within one time-step the heat is:

$$Q[x_n; s_{n-1} \to s_n] = E(x_n, s_n) - E(x_n, s_{n-1}) \ .$$

Over the course of driving the system from $t = 0$ to $t = N\Delta t = \tau$, the net energy change is then:

$$\begin{aligned}\Delta E &= E(x_N, s_N) - E(x_0, s_0) \\ &= W + Q \ ,\end{aligned}$$

where the net work and net heat are:

$$W = \sum_{n=1}^{N} W[x_{n-1} \to x_n; s_{n-1}]$$

and:

$$Q = \sum_{n=1}^{N} Q[x_n; s_{n-1} \to s_n] \ ,$$

respectively. Here, and later on, Δ applied to a quantity (besides t, since Δt is the duration of a single time-step) refers to its net change over the course of driving.

When the system strongly couples to a substrate with uncontrolled energy fluxes, steady-state dynamics are often established far from equilibrium, even when environmental parameters are held fixed. That is, for fixed driving $\ldots xxxxx \ldots$, the system settles down to a NESS with a distribution over observable system states given by the *nonequilibrium potential* $\phi(x,s)$:

$$\pi_x(s) = e^{-\phi(x,s)} \ . \qquad (3)$$

In this, $\phi(x,s)$ plays a role roughly analogous to energy eigenvalues. Thus, the thermodynamics of accessible energetics—the excess heat generated and work irretrievably performed in driving between NESSs—follows analogously to its equilibrium counterpart. This is complementary to recent SST studies [16–18, 23].

If steady-state free energies $F_{\mathrm{ss}}(x)$ and *effective* energies $E_{\mathrm{eff}}(x,s)$ could be uniquely (and usefully) defined, then the nonequilibrium potential would be:

$$\phi(x,s) = \beta[E_{\mathrm{eff}}(x,s) - F_{\mathrm{ss}}(x)] \ .$$

However, the assignment of steady-state free energies is problematic. Nevertheless, $\phi(x,s)$ retains meaning since it quantifies the steady-state *surprisal* of observing state s:

$$\phi(x,s) = -\ln \pi_x(s) \ .$$

The surprisal is Shannon's *self-information* [34]—the unaveraged individual-event entropy measuring how surprising a specific event is. Intuitively, we must do work to make otherwise unlikely things happen.

SST's *excess* work and heat can be defined via changes in steady-state surprisal ϕ, analogous to how equilibrium quantities are in terms of energy changes. For clarity, we temporarily restrict ourselves to the isothermal setting, but we can easily adapt to time-varying temperatures.

Excess work W_{ex} is environmentally driven change in nonequilibrium potential:

$$\begin{aligned}W_{\mathrm{ex}}&[x_{n-1} \to x_n; s_{n-1}] \\ &= \beta^{-1}[\phi(x_n, s_{n-1}) - \phi(x_{n-1}, s_{n-1})] ~,\end{aligned}$$

over one time-step. *Excess heat* Q_{ex} is the change in nonequilibrium potential due to the system's response:

$$Q_{\mathrm{ex}}[x_n; s_{n-1} \to s_n] = \beta^{-1}[\phi(x_n, s_n) - \phi(x_n, s_{n-1})] ~,$$

over one time-step. When driving from $t = 0$ to $t = N\Delta t = \tau$, the net change in nonequilibrium potential is:

$$\begin{aligned}\Delta\phi &= \phi(x_N, s_N) - \phi(x_0, s_0) \\ &= \beta(W_{\mathrm{ex}} + Q_{\mathrm{ex}}) \\ &= -\ln \frac{\pi_{x_N}(s_N)}{\pi_{x_0}(s_0)} ~, \end{aligned} \tag{4}$$

where the net excess work and net excess heat are:

$$W_{\mathrm{ex}} = \sum_{n=1}^{N} W_{\mathrm{ex}}[x_{n-1} \to x_n; s_{n-1}] \tag{5}$$

and:

$$Q_{\mathrm{ex}} = \sum_{n=1}^{N} Q_{\mathrm{ex}}[x_n; s_{n-1} \to s_n] ~, \tag{6}$$

respectively.

This approach to excess heat Q_{ex} coincides with SST's definition and reduces to total heat in equilibrium transitions. Importantly, it follows as closely as possible the equilibrium approach to total heat Q outlined above and deviates from the typical starting point: $Q_{\mathrm{ex}} \equiv Q - Q_{\mathrm{hk}}$, where Q_{hk} is the so-called *housekeeping heat* [35]. In contrast, excess work W_{ex} does *not* reduce to the total work in equilibrium transitions. Rather, W_{ex} goes over to $W - \Delta F_{\mathrm{eq}}$, if the steady states are near equilibrium. And, this fortuitously coincides with its previous narrower use in describing transitions atop equilibrium steady states—the work exerted beyond the change in free energy [36].

The energy $-Q_{\mathrm{ex}}$ that a system loses as excess heat can be interpreted as the heat dissipated due to relaxation during transitions between NESSs. Similarly, the excess work W_{ex} can be interpreted as the energy that *would* be dissipated if the system is allowed to relax back to a NESS. The difference between excess work W_{ex} and *dissipated work*, denoted W_{diss}, depends on a notion of excess nonequilibrium free energy, discussed shortly.

This framing reminds us that heat circumscribes how small, possibly intelligent, systems can store and transform energy via their *own agency.* For example, an increase in heat may indicate that a system has harvested energy, and the emission of heat may indicate an intrinsic computation [37] in the system's state-space. The *efficiency* of such tradeoffs—spending stored energy to achieve some utility—has been an active area of investigation recently, especially for small biological systems [38–40].

B. Excess environmental entropy production

In isothermal transitions between equilibrium steady states, the environmental entropy production is [24]:

$$\begin{aligned}\Omega_{\mathrm{eq}} &= \beta(W - \Delta F_{\mathrm{eq}}) \\ &= -\beta Q - \ln \frac{\pi_{x_N}(s_N)}{\pi_{x_0}(s_0)} ~.\end{aligned}$$

This extends to SST by defining the *excess environmental entropy production*:

$$\begin{aligned}\Omega &= \beta W_{\mathrm{ex}} \\ &= -\beta Q_{\mathrm{ex}} - \ln \frac{\pi_{x_N}(s_N)}{\pi_{x_0}(s_0)} ~, \end{aligned} \tag{7}$$

This has also been referred to as the "nonadiabatic component of entropy production" [18, 23, 41, 42]. Note that:

$$-\ln\left(\pi_{x_N}(s_N)/\pi_{x_0}(s_0)\right) = \Delta\phi ~,$$

recovering Eq. (4)'s change in nonequilibrium potential ϕ.

Recalling the definitions of Q_{ex} in terms of steady state surprisals and that $\phi(x, s) = -\ln \pi_x(s)$, we see that:

$$\begin{aligned}e^{-\beta Q_{\mathrm{ex}}} &= e^{-\sum_{n=1}^{N}[\phi(x_n,s_n)-\phi(x_n,s_{n-1})]} \\ &= \prod_{n=1}^{N} \frac{\pi_{x_n}(s_n)}{\pi_{x_n}(s_{n-1})} ~. \end{aligned} \tag{8}$$

And so, Eq. (7) gives:

$$\begin{aligned}e^{\Omega(x_{0:N+1}, s_{0:N})} &= \frac{\pi_{x_0}(s_0)}{\pi_{x_N}(s_N)} \prod_{n=1}^{N} \frac{\pi_{x_n}(s_n)}{\pi_{x_n}(s_{n-1})} \\ &= \prod_{n=0}^{N-1} \frac{\pi_{x_n}(s_n)}{\pi_{x_{n+1}}(s_n)} ~. \end{aligned} \tag{9}$$

If temperature varies, then the above still holds if we replace the steady-state probabilities with the temperature-dependent steady-state probabilities. Thus, to go beyond the isothermal setting, we use Eq. (9) as the defining relationship for the excess environmental entropy production Ω. If temperature is spatially homogeneous, then Ω is equivalent to:

$$\Omega = \Delta\phi - \tfrac{1}{k_\mathrm{B}} \int \frac{\delta Q_\mathrm{ex}}{T} ~.$$

However, spatially inhomogeneous temperatures can also be addressed by folding temperature dependence into the environmental input x.

We return to these expressions and explore their role in generalized fluctuation theorems once we develop the necessary quantitative notions of irreversibility in the following section. Immediately, though, we must address how the entropies of system paths determine free energies.

C. Path-induced entropies

In steady state, the system state probability distribution has a Boltzmann exponential dependence on the effective energies. Naturally, out of steady state the distribution is something different. There is a nonsteady-state free energy associated with this out-of-steady-state distribution, since the system can do work (or computations) at the cost of relaxing the distribution.

Nonsteady-state free energies are controlled by path-induced entropies, which come in several varieties. Here, we are especially interested in the controllable unaveraged state surprisals induced by the driving path $\overleftarrow{x}$:

$$h^{(s|x_{-\infty:t+1})} = -\ln \Pr(\mathcal{S}_t = s | x_{-\infty:t+1}) ~. \tag{10}$$

Since a semi-infinite history $\overleftarrow{x}$ induces a particular distribution over system states, this can be usefully recast in terms of the initial distribution $\boldsymbol{\mu}_0$ induced by the path $x_{-\infty:1}$ and the driving history $x_{1:t+1}$ since then:

$$\begin{aligned} h^{(s|\boldsymbol{\mu}_0, x_{1:t+1})} &= -\ln \Pr(\mathcal{S}_t = s | \mathcal{S}_0 \sim \boldsymbol{\mu}_0, x_{1:t+1}) \\ &= -\ln \langle \boldsymbol{\mu}_0 | \, \mathsf{T}^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}} | x_{1:t+1})} \, | s \rangle ~, \end{aligned} \tag{11}$$

where $\Pr(\mathcal{S}_t = s | \mathcal{S}_0 \sim \boldsymbol{\mu}_0, x_{1:t+1})$ is the probability that the state is s at time t, under the measure induced when the initial state $\mathcal{S}_0 \sim \boldsymbol{\mu}_0$ (distributed according to $\boldsymbol{\mu}_0$) [43] and given the driving history $x_{1:t+1} = x_1 \ldots x_t$ since the initial time.

Alternatively, consider the distribution $\boldsymbol{\mu}$ induced from a start distribution by the driving history since the start. Then the path-induced state-surprisal can be expressed simply in terms of the *present* environmental-history-induced distribution over system states and the candidate state s:

$$\begin{aligned} h^{(s|\boldsymbol{\mu})} &= -\ln \Pr(\mathcal{S}_t = s | \mathcal{S}_t \sim \boldsymbol{\mu}) \\ &= -\ln \langle \boldsymbol{\mu} | s \rangle ~. \end{aligned} \tag{12}$$

Thermodynamic units of entropy are recovered by multiplying the Shannon-like path surprisals by Boltzmann's constant: $\mathfrak{s} = k_\mathrm{B} h$.

Averaging the path-induced state-surprisal over states gives a genuine input-conditioned Shannon entropy:

$$\begin{aligned} \langle h^{(s_t|\overleftarrow{x}_t)} \rangle_{\Pr(s_t|\overleftarrow{x}_t)} &= -\Big\langle \ln \Pr(s_t | \overleftarrow{x}_t) \Big\rangle_{\Pr(s_t|\overleftarrow{x}_t)} \\ &= -\sum_{s_t} \Pr(s_t|\overleftarrow{x}_t) \ln \Pr(s_t|\overleftarrow{x}_t) \\ &= \mathrm{H}[\mathcal{S}_t | \overleftarrow{X}_t = \overleftarrow{x}_t] ~, \end{aligned} \tag{13}$$

where $\mathrm{H}[\cdot|\cdot]$ is the conditional Shannon entropy in units of nats.

It follows directly that the state-averaged path-induced entropy $k_\mathrm{B} \, \mathrm{H}[\mathcal{S}_t | \overleftarrow{x}_t]$ is an extension of the system's steady-state nonequilibrium entropy S_ss. That is, in steady-state the state-averaged path-induced entropy reduces to:

$$\begin{aligned} k_\mathrm{B} \, \mathrm{H}[\mathcal{S}_t | \overleftarrow{X}_t = \ldots xxx] &= -k_\mathrm{B} \, \mathrm{H}[\mathcal{S}_t | \mathcal{S}_t \sim \boldsymbol{\pi_x}] \\ &= -k_\mathrm{B} \sum_{s \in \boldsymbol{\mathcal{S}}} \pi_x(s) \ln \pi_x(s) \\ &= S_\mathrm{ss}(x) ~. \end{aligned} \tag{14}$$

The system steady-state nonequilibrium entropy S_ss has been discussed as a fundamental entity in SST; e.g., see Refs. [16] and [18]. However, Eq. (13) ($\times k_\mathrm{B}$) gives the appropriate extension for the thermodynamic entropy of a nonequilibrium system that is *not in steady state.* Rather, it is the entropy over system states given the entire history of environmental driving.

When $\boldsymbol{\mathcal{S}}$ is the set of microstates, rather than, say, observable mesoscopic states, the unaveraged nonequilibrium free energy F enjoys the familiar relationship between energy E and (path-induced) entropy $\mathfrak{s}$:

$$\begin{align} F^{(s_t|x_{-\infty:t+1})} &\equiv E(x_t, s_t) - T \mathfrak{s}^{(s_t|x_{-\infty:t+1})} \tag{15} \\ &= F_\mathrm{eq}(x_t) + \beta^{-1} \ln \frac{\Pr(s_t|x_{-\infty:t+1})}{\pi_{x_t}(s_t)} ~. \tag{16} \end{align}$$

Or, averaging over states:

$$\begin{aligned} \mathcal{F}(t) &= U(t) - \beta^{-1} \, \mathrm{H}[\mathcal{S}_t | x_{-\infty:t+1}] \\ &= F_\mathrm{eq}(x_t) + \beta^{-1} D_\mathrm{KL}\big(\Pr(\mathcal{S}_t|x_{-\infty:t+1}) \,||\, \boldsymbol{\pi}_{x_t}\big) ~, \end{aligned} \tag{17}$$

where $\mathcal{F}(t)$ is the expected instantaneous nonequilibrium free energy, $U(t)$ is the expected instantaneous thermal energy, and $D_{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback–Leibler divergence [34]. Recognizing $k_{\mathrm{B}}\,\mathrm{H}[\mathcal{S}_t|x_{-\infty:t+1}]$ as the natural extension of a system's thermodynamic entropy, Eq. (17) is familiar from equilibrium thermodynamics, but it is now applicable arbitrarily far from equilibrium and at any time t using the instantaneous temperature. This is not the first statement of such a generalized relationship; compare, e.g., Refs. [44–46]. In equilibrium, the expected value of the path-induced entropy (using microstates) reduces to *the* equilibrium entropy of a system.

In the setting of effective states and NESS surprisals, though, we can no longer directly use Eq. (15). Nevertheless, by analogy with Eq. (16), we can still identify the *nonsteady-state addition* $\gamma(\cdot|\cdot)$ *to free energy* as:

$$\beta^{-1}\gamma(s|\boldsymbol{\mu},x) \equiv \beta^{-1}\ln\frac{\Pr(\mathcal{S}_t = s|\mathcal{S}_{t-1}\sim\boldsymbol{\mu}, X_t = x)}{\pi_x(s)} \ . \tag{18}$$

Expressed differently, it is:

$$\begin{aligned}\gamma(s|\boldsymbol{\mu},x) &= h^{(s|\boldsymbol{\pi_x})} - h^{(s|\boldsymbol{\mu},x)} \\ &= \phi(x,s) - h^{(s|\boldsymbol{\mu},x)} \ .\end{aligned}$$

Averaging over states this becomes the Kullback–Leibler divergence between nonsteady state and steady state distributions:

$$\langle\gamma(s|\boldsymbol{\mu},x)\rangle = D_{\mathrm{KL}}\left[\Pr(\mathcal{S}_t|\mathcal{S}_{t-1}\sim\boldsymbol{\mu}, X_t = x)\,||\,\boldsymbol{\pi_x}\right] \ ,$$

which is nonnegative.

Identifying the nonsteady-state contribution to the free energy allows us to introduce the *dissipated work*:

$$\begin{aligned}W_{\mathrm{diss}} &\equiv W_{\mathrm{ex}} - \beta^{-1}\Delta\gamma & (19)\\ &= -Q_{\mathrm{ex}} + \beta^{-1}\Delta h \ , & (20)\end{aligned}$$

to account for the fact that excess work is not fully dissipated until the distribution relaxes back to steady state $\boldsymbol{\pi_x}$. An important consequence is that the excess work not *yet* dissipated can be reclaimed—by a clever control protocol that induces a subsequent "fluctuation" with $W_{\mathrm{ex}} < 0$ in the midst of a driven nonequilibrium excursion.

The role of the nonsteady-state contribution to free energy will be apparent in the FTs to come shortly. This generalizes similar FTs that are restricted to starting and possibly ending in a steady state $\boldsymbol{\pi_x}$. The generalization here is key to analyzing complex systems, since many simply cannot be initiated in a steady state without losing their essential character.

IV. IRREVERSIBILITY

To emphasize, the preceding did not reference and does not require detailed balance. However, to ground the coming development, we need to describe the roles of reversibility, detailed balance, and their violations. At a minimum, this is due to most FTs assuming reversibility of the effective dynamic over states. Having established the necessary concepts and giving a measure of the irreversibility of the effective dynamic, we finally move on to FTs for nondetailed-balanced processes.

A. Detailed balance

Transitioning from state a to state b, say, invoking detailed balance assumes that:

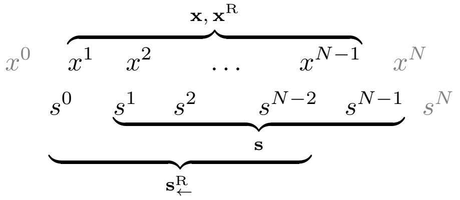
$$\frac{\Pr(\mathcal{S}_n = a|\mathcal{S}_{n-1} = b, X_n = x)}{\Pr(\mathcal{S}_n = b|\mathcal{S}_{n-1} = a, X_n = x)} = \frac{\pi_x(a)}{\pi_x(b)} \ .$$

Though, we do *not* assume detailed balance over the states considered here, we refer to it occasionally. For example, assuming detailed balance, microscopic reversibility and the standard Crooks fluctuation theorem (CFT) follow almost immediately.

In contrast, complex systems sustained out of equilibrium by an active substrate generically evolve via nondetailed-balanced dynamics. To wit, many examples of nondetailed-balanced dynamics are exhibited by chemical kinetics in biological systems [47–49]. We conclude with a thermodynamic analysis of one neurobiological example.

B. Microscopic reversibility

Consider a particular realization of interleaved environmental-input sequence $\ldots x^1 x^2 \cdots x^{N-1}\ldots$ and system-state sequence $\ldots s^1 s^2 \cdots s^{N-1}\ldots$:



There are several length-$(N-1)$ subsequences in play here, including the forward trajectory $\mathbf{x} = x^1 x^2 \cdots x^{N-2} x^{N-1}$ of the environmental driving and the forward trajectory $\mathbf{s} = s^1 s^2 \ldots s^{N-2} s^{N-1}$ of the state sequence. Furthermore, let $\mathbf{x}^{\mathrm{R}} = x^{N-1} x^{N-2} \ldots x^2 x^1$ be the time-reversal of the environmental driving $\mathbf{x}$ and

$\mathbf{s}^{\mathrm{R}}_{\leftarrow} = s^{N-2}s^{N-3}\ldots s^1 s^0$ the time-reversal of the time-shifted state sequence $\mathbf{s}$.

For example, if $\mathcal{X} = \{0,1\}$ and $\boldsymbol{\mathcal{S}} = \{a,b,c\}$, then $\mathbf{x}$ may be the sequence $00101110\ldots11000010$ and $\mathbf{s}$ the sequence $acaaaaba\ldots abaccabc$. Then $\mathbf{x}^{\mathrm{R}}$ is the sequence $01000011\ldots01110100$. Taking the time reversal of the state sequence, we have $cbaccaba\ldots abaaaaca$. However, since $\mathbf{s}^{\mathrm{R}}_{\leftarrow}$ is also time-shifted by one time-step, we must drop the first c and append another symbol, say a. Then $\mathbf{s}^{\mathrm{R}}_{\leftarrow}$ is the sequence $baccaba\ldots abaaaacaa$.

Let Q_{F} be the excess heat of the joint forward sequences $\mathbf{x}$ and $s^0\mathbf{s}$, according to Eq. (6). By definition, a system–environment effective dynamic is *microscopically reversible* if:

$$\frac{\Pr(\mathcal{S}_{1:N} = \mathbf{s}|\mathcal{S}_0 = s^0, X_{1:N} = \mathbf{x})}{\Pr(\mathcal{S}_{1:N} = \mathbf{s}^{\mathrm{R}}_{\leftarrow}|\mathcal{S}_0 = s^{N-1}, X_{1:N} = \mathbf{x}^{\mathrm{R}})} = e^{-\beta Q_{\mathrm{F}}} ~,$$

for any $s^0 \in \boldsymbol{\mathcal{S}}$, $\mathbf{s} \in \boldsymbol{\mathcal{S}}^{N-1}$, and $\mathbf{x} \in \mathcal{X}^{N-1}$. As a useful visual aid, we can re-express this as:

$$\frac{\Pr(s^0 \xrightarrow{x^1} s^1 \cdots s^{N-2} \xrightarrow{x^{N-1}} s^{N-1}|s^0, \mathbf{x})}{\Pr(s^0 \xleftarrow{x^1} s^1 \cdots s^{N-2} \xleftarrow{x^{N-1}} s^{N-1}|s^{N-1}, \mathbf{x}^{\mathrm{R}})} = e^{-\beta Q_{\mathrm{F}}} ~.$$

Otherwise, microscopic reversibility is broken.

Although, microscopic reversibility has also been referred to as a "detailed fluctuation theorem", it is actually an assumption appropriate only in special cases. For example, Eq. (8) shows that if the dynamics are Markovian over states (given input) and obey detailed balance (à la §IV A), then microscopic reversibility is satisfied for arbitrary non-Markovian inputs. In essence, this is the justification of microscopic reversibility suggested by Crooks [2, 24] from which his eponymous fluctuation theorem follows.

In this view, detailed balance and microscopic reversibility are effectively the same assumption since each implies the other. Section § V C generalizes the Crooks Fluctuation Theorem to describe fluctuation laws in the absence of microscopic reversibility.

C. Path dependence and irreversibility

The importance of state-space path dependence is captured via an informational quantity Υ we call the *path relevance* of a state sequence $s_{1:N}$ given initial state s_0 and input sequence $x_{1:N}$:

$$\begin{aligned}\Upsilon(s_{1:N}|s_0, x_{1:N}) &\equiv \ln \frac{\Pr(s_{1:N}|s_0, x_{1:N})}{\prod_{n=1}^{N-1} \pi_{x_n}(s_n)} \\ &= \sum_{n=1}^{N-1} \ln \frac{\Pr(s_n|s_{0:n}, x_{1:N})}{\pi_{x_n}(s_n)} ~.\end{aligned} \qquad (21)$$

(The branching Pythagorean letter Υ recognizes its ancient symbolism—divergent consequences of choosing one path over another.) Note that the steady-state probabilities in the denominator are independent of the ordering of the state sequence, whereas the numerator (even after factoring) depends on the probabilities of state-to-state transitions. Hence, Υ quantifies the log probability of the driving-induced state transitions, discounted by the asymptotic probabilities of state occupation.

A joint sequence lacks path relevance, if $\Upsilon = 0$ for that sequence. However, $\Upsilon = 0$ is only typical of the rather structureless collection of independent and identically distributed (IID) processes.

Whenever state transitions are Markovian given the input, the numerator in Eq. (21) simplifies to:

$$\Pr(s_{1:N}|x_{1:N}, s_0) = \prod_{n=1}^{N-1} \Pr(s_n|s_{n-1}, x_n) ~,$$

and the path relevance becomes:

$$\Upsilon = \sum_{n=1}^{N-1} \ln \frac{\Pr(s_n|s_{n-1}, x_n)}{\pi_{x_n}(s_n)} ~.$$

Equation (18) shows that this is the accumulated nonequilibrium additions to free energy when the system, known to be in a certain state, jumps to a new state under the driving influence. Thus, there is path relevance even for Markov processes. The actual driving history matters. When a system is *non*-Markovian, there are yet additional contributions to path relevance.

Path relevance of a particular state sequence given a particular driving is a system feature, regardless of the environment in which the system finds itself. However, expectation values involving the above relationship can reflect the environment's nature.

For our development, we find it useful to consider both the forward-path dependence and the reverse-path dependence of a particular joint sequence: $x^1 \ldots x^{N-1}$ and $s^0 \ldots s^{N-1}$. The *forward-path dependence* is as expected:

$$\begin{aligned}\Upsilon &= \Upsilon(\mathbf{s}|s^0, \mathbf{x}) \\ &= \ln \frac{\Pr(\mathbf{s}|s^0, \mathbf{x})}{\prod_{n=1}^{N-1} \pi_{x^n}(s^n)} ~,\end{aligned} \qquad (22)$$

and, similarly, the *reverse-path dependence* $⅄$ is:

$$\begin{aligned} ⅄ &= ⅄(\mathbf{s}|s^0, \mathbf{x}) \\ &= \Upsilon(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1}, \mathbf{x}^{\mathrm{R}}) \\ &= \ln \frac{\Pr(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1}, \mathbf{x}^{\mathrm{R}})}{\prod_{n=1}^{N-1} \pi_{x^n}(s^{n-1})} ~. \end{aligned} \tag{23}$$

The reverse-path dependence quantifies the log-probability that the reverse driving induces the reverse state-sequence, discounted by the state-path-independent asymptotic probabilities of dwelling in each state.

And, finally, we have the *net directional relevance*—of a particular path $\mathbf{s}$ given s^0 and $\mathbf{x}$—the *irreversibility*:

$$\Psi \equiv \Upsilon - ⅄ ~. \tag{24}$$

Nonzero Ψ indicates the irrevocable consequences of path traversal. *Microscopically reversible dynamics have* $\Psi = 0$ *for all paths with nonzero probability*, indicating no divergence in path branching anywhere through the state-space. And so, $\Psi = 0$ for all paths with nonzero probability if and only if the dynamic satisfies detailed balance. In short, Ψ quantifies the imbalance in path reciprocity along a driven state-sequence.

Sometimes $⅄$ can be $-\infty$ for an allowed forward path $\Pr(\mathbf{s}|s^0, \mathbf{x}) > 0$, corresponding to a forbidden reverse path $\Pr(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1}, \mathbf{x}^{\mathrm{R}}) = 0$. This is a situation that never arises with detailed balance dynamics. Such paths are infinitely irreversible: $\Psi = \infty$.

V. GENERALIZED FLUCTUATION THEOREMS FOR NONEQUILIBRIUM SYSTEMS

Absent microscopic reversibility, the architecture of transitions over state-space matters. More concretely, we will constructively show how this architecture affects the nonequilibrium thermodynamics of complex systems.

A. Generalized detailed fluctuation theorem

Assume the system under study starts from some distribution $\boldsymbol{\mu}_{\mathrm{F}}$ and that the associated reverse trajectory (when starting from some other distribution $\boldsymbol{\mu}_{\mathrm{R}}$) is allowed—that is, it has nonzero probability. Then the ratio of conditional probabilities of a state sequence (given a driving sequence) to the reversed state sequence (given reversed driving) is:

$$\begin{aligned} &\frac{\Pr(\boldsymbol{\mu}_{\mathrm{F}} \xrightarrow{x^0} s^0 \xrightarrow{x^1} s^1 \cdots s^{N-2} \xrightarrow{x^{N-1}} s^{N-1}|\boldsymbol{\mu}_{\mathrm{F}}, x^0\mathbf{x})}{\Pr(s^0 \xleftarrow{x^1} s^1 \cdots s^{N-2} \xleftarrow{x^{N-1}} s^{N-1} \xleftarrow{x^N} \boldsymbol{\mu}_{\mathrm{R}}|\boldsymbol{\mu}_{\mathrm{R}}, x^N\mathbf{x}^{\mathrm{R}})} \\ &\qquad = \frac{\Pr(\mathcal{S}_{0:N} = s^0\mathbf{s}|\mathcal{S}_{-1} \sim \boldsymbol{\mu}_{\mathrm{F}}, X_{0:N} = x^0\mathbf{x})}{\Pr(\mathcal{S}_{0:N} = s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|\mathcal{S}_{-1} \sim \boldsymbol{\mu}_{\mathrm{R}}, X_{0:N} = x^N\mathbf{x}^{\mathrm{R}})} \\ &\qquad = \frac{\Pr(s^0|\boldsymbol{\mu}_{\mathrm{F}}, x^0)}{\Pr(s^{N-1}|\boldsymbol{\mu}_{\mathrm{R}}, x^N)} \; \frac{\Pr(\mathbf{s}|s^0, \mathbf{x})}{\Pr(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1}, \mathbf{x}^{\mathrm{R}})} \\ &\qquad = \frac{\Pr(s^0|\boldsymbol{\mu}_{\mathrm{F}}, x^0)}{\pi_{x^0}(s^0)} \; \frac{\pi_{x^N}(s^{N-1})}{\Pr(s^{N-1}|\boldsymbol{\mu}_{\mathrm{R}}, x^N)} \; \frac{\Pr(\mathbf{s}|s^0, \mathbf{x})}{\prod_{n=1}^{N-1} \pi_{x^n}(s^n)} \; \frac{\prod_{n=1}^{N-1} \pi_{x^n}(s^{n-1})}{\Pr(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1}, \mathbf{x}^{\mathrm{R}})} \; \prod_{n=0}^{N-1} \frac{\pi_{x^n}(s^n)}{\pi_{x^{n+1}}(s^n)} \\ &\qquad = \frac{\Pr(s^0|\boldsymbol{\mu}_{\mathrm{F}}, x^0)}{\pi_{x^0}(s^0)} \; \frac{\pi_{x^N}(s^{N-1})}{\Pr(s^{N-1}|\boldsymbol{\mu}_{\mathrm{R}}, x^N)} \; e^{\Psi_{\mathrm{F}}} \, e^{\Omega_{\mathrm{F}}} \\ &\qquad = e^{\gamma(s^0|\boldsymbol{\mu}_{\mathrm{F}}, x^0) - \gamma(s^{N-1}|\boldsymbol{\mu}_{\mathrm{R}}, x^N)} \; e^{\Omega_{\mathrm{F}} + \Psi_{\mathrm{F}}} ~, \end{aligned} \tag{25}$$

where $\Omega_{\mathrm{F}} = \Omega(X_{0:N+1} = x^0\mathbf{x}x^N, \mathcal{S}_{0:N} = s^0\mathbf{s})$ is the excess environmental entropy production in the forward trajectory, $\Psi_{\mathrm{F}} = \Psi(\mathcal{S}_{0:N} = s^0\mathbf{s}|X_{1:N} = \mathbf{x})$ is the irreversibility of the forward trajectory, and $\gamma(s|\boldsymbol{\mu}, x) = \ln\left(\Pr(s|\boldsymbol{\mu}, x)/\pi_x(s)\right)$ is the nonsteady-state addition to free energy associated with being in the nonsteady-state distribution $\boldsymbol{\mu}$ with environmental drive x.

Since Ψ_{F} can diverge for forward paths with nonzero probability, we typically rewrite Eq. (25) as the "less divergent" expression:

$$\begin{aligned} &e^{-\gamma_{\mathrm{F}}} \Pr(s^0\mathbf{s}|\boldsymbol{\mu}_{\mathrm{F}}, x^0\mathbf{x}) e^{-\Psi_{\mathrm{F}}} \\ &\qquad = e^{-\gamma_{\mathrm{R}}} \Pr(s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|\boldsymbol{\mu}_{\mathrm{R}}, x^N\mathbf{x}^{\mathrm{R}}) e^{\Omega_{\mathrm{F}}} ~. \end{aligned} \tag{26}$$

Eq. (26) is the fundamental relation for all that follows: it relates the probabilities of forward and reverse trajectories via entropy production Ω_{F} of the forward path, irreversibility Ψ_{F} of the forward path, and change $\beta^{-1}(\gamma_{\mathrm{F}} - \gamma_{\mathrm{R}})$ in the nonsteady-state addition to free energy between the forward and reverse start-distributions.

In what follows it will be all too easy to write seemingly divergent expressions. Such divergences do not manifest themselves when taking expectation values for physical quantities involving them, since they come weighted with zero probability. This is similar to the reasonable convention for Shannon entropies that $0 \log 0 = 0$. Nevertheless, caution is advised when probabilities vanish.

B. Simplifications

Before proceeding and to aid understanding, let's consider several special cases. If the forward drive or protocol begins with the system in the steady state induced by the static environmental drive x^0, then $\boldsymbol{\mu}_{\mathrm{F}} = \boldsymbol{\pi}_{\boldsymbol{x}^0}$ and $\gamma(s^0|\boldsymbol{\mu}_{\mathrm{F}}, x^0) = 0$. Similarly, if the reverse protocol begins with the system in the steady state induced by the static environmental drive x^N, then $\boldsymbol{\mu}_{\mathrm{R}} = \boldsymbol{\pi}_{\boldsymbol{x}^N}$ and $\gamma(s^N|\boldsymbol{\mu}_{\mathrm{R}}, x^N) = 0$. In this case, Eq. (26) simplifies to:

$$\frac{\Pr(\boldsymbol{\pi}_{\boldsymbol{x}^0} \xrightarrow{x^0} s^0 \cdots \xrightarrow{x^{N-1}} s^{N-1}|\boldsymbol{\pi}_{\boldsymbol{x}^0}, x^0\mathbf{x})}{\Pr(s^0 \xleftarrow{x^1} \cdots s^{N-1} \xleftarrow{x^N} \boldsymbol{\pi}_{\boldsymbol{x}^N}|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}})} = e^{\Omega_{\mathrm{F}}+\Psi_{\mathrm{F}}} \ .$$

As a separate matter, if the dynamics are microscopically reversible, then $\Psi = 0$. Consider the very special case where (i) the dynamics are microscopically reversible, (ii) the forward driving begins with the system equilibrated with x^0, and (iii) the reverse driving begins with the system equilibrated with x^N. Then, the ratio of probabilities of observing a forward state sequence (given forward driving) and observing the reversal of that state sequence (given the reversal of that driving) is simply $e^{\Omega_{\mathrm{F}}}$. That is, the forward sequence is exponentially more likely if it has positive entropy production.

Apparently, the more general case is more nuanced and, beyond depending on a nonsteady-state starting distribution, it depends strongly on the architecture of branching transitions among states.

Another interesting special case is if $x^N = x^{N-1}$ and $\boldsymbol{\mu}_{\mathrm{R}}$ is the distribution $\boldsymbol{\mu}(\boldsymbol{\mu}_{\mathrm{F}}, \mathbf{x})$ that the forward driving induces from $\boldsymbol{\mu}_{\mathrm{F}}$. Then the *dissipated work* $W_{\mathrm{diss}} \equiv W_{\mathrm{ex}} - \beta^{-1}\Delta\gamma$ associated with the forward trajectory comes into play. (Recall that $\beta^{-1}\Delta\gamma$ is the change in nonsteady-state contributions to free energy.) Then the ratio of forward- and reverse-path probabilities is:

$$\begin{aligned}\frac{\Pr(\boldsymbol{\mu}_{\mathrm{F}} \xrightarrow{x^0} \cdots \xrightarrow{x^{N-1}} s^{N-1}|\boldsymbol{\mu}_{\mathrm{F}}, \mathbf{x})}{\Pr(s^0 \xleftarrow{x^1} \cdots \xleftarrow{x^N} \boldsymbol{\mu}(\boldsymbol{\mu}_{\mathrm{F}}, \mathbf{x})|\boldsymbol{\mu}(\boldsymbol{\mu}_{\mathrm{F}}, \mathbf{x}), \mathbf{x}^{\mathrm{R}})} & \\ = e^{\Psi_{\mathrm{F}}} e^{\beta[W_{\mathrm{ex}} - \beta^{-1}\Delta\gamma]} & \\ = e^{\Psi_{\mathrm{F}}} e^{\beta W_{\mathrm{diss}}} \ . & \qquad (27)\end{aligned}$$

Even in the case of microscopic reversibility, this is useful since it generalizes previous FTs to nonequilibrium start and end distributions. In the case of microscopic reversibility, $\Psi_{\mathrm{F}} = 0$ and so the ratio of forward- and reverse-path probabilities from any nonequilibrium start and end distribution is exponential $e^{\beta W_{\mathrm{diss}}}$ in the dissipated work. Thus, an experimental test of this result is one with time-symmetric driving. The forward protocol corresponds to the first half of the driving while the reverse protocol is the second half. Clearly, the final nonequilibrium distribution for the forward protocol is the same as the initial nonequilibrium distribution for the reverse protocol. The dissipated work then corresponds to that dissipated in the first half of the driving. Practically, in cases where the dynamic is not microscopically reversible, this allows experimentally extracting the system's irreversibility Ψ.

C. Generalized Crooks fluctuation Theorem

We can now turn to the irreversible analog of the Crooks Fluctuation Theorem (CFT).

First, we note that both entropy production Ω and irreversibility Ψ are odd under time reversal. Explicitly, we have:

$$\begin{aligned}\Omega(X_{0:N+1} = x^0\mathbf{x}x^N, S_{0:N} = s^0\mathbf{s}) & \\ = \ln \prod_{n=0}^{N-1} \frac{\pi_{x^n}(s^n)}{\pi_{x^{n+1}}(s^n)} & \\ = -\ln \prod_{n=0}^{N-1} \frac{\pi_{x^{n+1}}(s^n)}{\pi_{x^n}(s^n)} & \\ = -\ln \prod_{n=0}^{N-1} \frac{\pi_{x^{N-n}}(s^{N-1-n})}{\pi_{x^{N-1-n}}(s^{N-1-n})} & \\ = -\Omega(X_{0:N+1} = x^N\mathbf{x}^{\mathrm{R}}x^0, S_{0:N} = s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}) &\end{aligned}$$

and:

$$\begin{aligned}
&\Psi(S_{0:N} = s^0\mathbf{s}|X_{1:N} = \mathbf{x}) \\
&\quad = \ln\Big[\frac{\Pr(\mathbf{s}|s^0,\mathbf{x})}{\Pr(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1},\mathbf{x}^{\mathrm{R}})}\prod_{n=1}^{N-1}\frac{\pi_{x^n}(s^{n-1})}{\pi_{x^n}(s^n)}\Big] \\
&\quad = -\ln\Big[\frac{\Pr(\mathbf{s}^{\mathrm{R}}_{\leftarrow}|s^{N-1},\mathbf{x}^{\mathrm{R}})}{\Pr(\mathbf{s}|s^0,\mathbf{x})}\prod_{n=1}^{N-1}\frac{\pi_{x^n}(s^n)}{\pi_{x^n}(s^{n-1})}\Big] \\
&\quad = -\Psi(S_{0:N} = s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|X_{1:N} = \mathbf{x}^{\mathrm{R}}) \ .
\end{aligned}$$

For brevity, let $\Omega_{\mathrm{F}} \equiv \Omega(X_{0:N+1} = x^0\mathbf{x}x^N, \mathcal{S}_{0:N} = s^0\mathbf{s})$ and $\Omega_{\mathrm{R}} \equiv \Omega(X_{0:N+1} = x^N\mathbf{x}^{\mathrm{R}}x^0, \mathcal{S}_{0:N} = s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow})$. And, similarly, $\Psi_{\mathrm{F}} \equiv \Psi(\mathcal{S}_{0:N} = s^0\mathbf{s}|X_{1:N} = \mathbf{x})$ and $\Psi_{\mathrm{R}} \equiv \Psi(\mathcal{S}_{0:N} = s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|X_{1:N} = \mathbf{x}^{\mathrm{R}})$. In this notation, we just established that $\Omega_{\mathrm{F}} = -\Omega_{\mathrm{R}}$ and $\Psi_{\mathrm{F}} = -\Psi_{\mathrm{R}}$.

Second, if we now choose $\boldsymbol{\mu}_{\mathrm{F}} = \boldsymbol{\pi}_{\boldsymbol{x}^0}$ and $\boldsymbol{\mu}_{\mathrm{R}} = \boldsymbol{\pi}_{\boldsymbol{x}^N}$ and marginalize over all possible state trajectories, we find that the joint probability of entropy production and irreversibility given the driving protocol starting from a steady-state distribution is:

$$\begin{aligned}
&\Pr(\Omega,\Psi|\boldsymbol{\pi}_{\boldsymbol{x}^0}, x^0\mathbf{x}x^N) \\
&\quad = \sum_{s^0\mathbf{s}\in\boldsymbol{\mathcal{S}}^N} \Pr(s^0\mathbf{s}|\boldsymbol{\pi}_{\boldsymbol{x}^0}, x^0\mathbf{x})\,\delta_{\Omega,\Omega_{\mathrm{F}}}\delta_{\Psi,\Psi_{\mathrm{F}}} \\
&\quad = \sum_{s^0\mathbf{s}\in\boldsymbol{\mathcal{S}}^N} e^{\Omega_{\mathrm{F}}}e^{\Psi_{\mathrm{F}}}\Pr(s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}})\,\delta_{\Omega,\Omega_{\mathrm{F}}}\delta_{\Psi,\Psi_{\mathrm{F}}} \\
&\quad = e^{\Omega}e^{\Psi}\sum_{s^0\mathbf{s}\in\boldsymbol{\mathcal{S}}^N} \Pr(s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}})\,\delta_{\Omega,\Omega_{\mathrm{F}}}\delta_{\Psi,\Psi_{\mathrm{F}}} \\
&\quad = e^{\Omega}e^{\Psi}\sum_{s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}\in\boldsymbol{\mathcal{S}}^N} \Pr(s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}})\,\delta_{\Omega,-\Omega_{\mathrm{R}}}\delta_{\Psi,-\Psi_{\mathrm{R}}} \\
&\quad = e^{\Omega}e^{\Psi}\Pr(-\Omega,-\Psi|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}}x^0) \ .
\end{aligned}$$

Finally, we rewrite this to give the *extended CFT for irreversible processes*:

$$\frac{\Pr(\Omega,\Psi|\boldsymbol{\pi}_{\boldsymbol{x}^0}, x^0\mathbf{x}x^N)}{\Pr(-\Omega,-\Psi|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}}x^0)} = e^{\Psi}e^{\Omega} \ . \qquad (28)$$

D. Interpretation

In the special case of isothermal time-symmetric driving—$x^0\mathbf{x}x^0 = x^0\mathbf{x}^{\mathrm{R}}x^0 = x^0x^1x^2\ldots x^2x^1x^0$—and starting from a steady-state distribution, Eq. (28) provides a useful comparison between values of excess work achieved by the single time-symmetric driving protocol:

$$\frac{\Pr(W_{\mathrm{ex}},\Psi)}{\Pr(-W_{\mathrm{ex}},-\Psi)} = e^{\Psi}e^{\beta W_{\mathrm{ex}}} \ . \qquad (29)$$

Equation (28) should be compared to the original CFT that, in its most general form, can be written (with necessary interpretation) as [50]:

$$\frac{\Pr_{\mathrm{F}}(\Omega)}{\Pr_{\mathrm{R}}(-\Omega)} = e^{\Omega} \ . \qquad (30)$$

It is tempting to write Eq. (30) as:

$$\frac{\Pr(\Omega|\boldsymbol{\pi}_{\boldsymbol{x}^0}, x^0\mathbf{x}x^N)}{\Pr(-\Omega|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}}x^0)} \overset{?}{=} e^{\Omega} \ . \qquad (31)$$

This form presents some concerns, however. In the case of detailed balance, though, $\Psi = 0$ for all trajectories, and so our Eq. (28) guarantees Eq. (31) in the case of detailed balance. Crooks' original CFT derivation [2, 24] also assumed detailed balance, and so Eq. (31) was implied.

However, absent detailed balance, Eq. (30) has a rather different interpretation: $\Pr_{\mathrm{R}}(\cdot)$ then implies not only the reversed driving, but also that the distribution describes a different "reversed" system that is not of direct physical relevance [50, 51]. One consequence is that the probabilities in the numerator and denominator are not comparable in any physical sense. So, in general, we have:

$$\frac{\Pr(\Omega|\boldsymbol{\pi}_{\boldsymbol{x}^0}, x^0\mathbf{x}x^N)}{\Pr(-\Omega|\boldsymbol{\pi}_{\boldsymbol{x}^N}, x^N\mathbf{x}^{\mathrm{R}}x^0)} \neq e^{\Omega} \ . \qquad (32)$$

In contrast, our irreversible CFT in Eq. (28) compares probabilities of excess entropy production (and path irreversibility) for the same thermodynamic system under a control protocol and under the reversed control protocol. Equation (28), unlike equalities involving an unphysical dual dynamic as in Eq. (30), allows a clear and meaningful physical interpretation of the relationship between entropies produced and, moreover, is not limited by assuming detailed balance.

Note that our Eq. (28), expressed in terms of excess environmental entropy production Ω and path irreversibility Ψ, does not make explicit mention of temperature. Indeed, if temperature dependence is folded into different environmental inputs x, then Eq. (28) applies just as well to systems driven by environments with spatially inhomogeneous temperature distributions that change in time. Explicitly, $\boldsymbol{\pi}_{\boldsymbol{x}}$ and $\boldsymbol{\pi}_{\boldsymbol{x}'}$ could represent the distribution over effective states induced by environmental conditions associated with x and x' *including* their different spatial distributions of temperature.

E. Translation to steady-state thermodynamics

A better understanding of the irreversible CFT comes by comparing it to recent related work. Most directly, our results complement those on driven transitions between

NESSs. Specifically, the importance of nondetailed-balanced dynamics in enabling the organization of complex nonequilibrium behavior has been considered previously. See for example the discussion in Ref. [44], which also introduced a path-induced entropy which is an ensemble average of that considered here.

Another comparison is found in Ref. [16]'s nonequilibrium thermodynamics over NESSs using housekeeping Q_{hk} and excess Q_{ex} heats. While that treatment focused on Langevin dynamics, we find that in general Q_{hk} corresponds directly to our path irreversibility Ψ. Specifically, in the isothermal setting there, according to Eq. (35), we have:

$$\beta Q_{\mathrm{hk}} \approx \Psi \ .$$

Indeed, for isothermal Markovian dynamics Eq. (7.7) of Ref. [52] suggests (via their Eqs. (2.11) and (7.1)) that this is in fact an equality:

$$\beta Q_{\mathrm{hk}} = \Psi \ . \tag{33}$$

Reference [23] called the irreversibility Ψ the *adiabatic contribution* to entropy production. Several related translations from Ref. [16] to our setting can also be easily made: $\rho_{\mathrm{ss}}(s;x) \to \pi_x(s)$, $\phi(s;x) \to -\log \pi_x(s)$, $\Delta S \to \Delta S_{\mathrm{ss}}$, and $\beta Q_{\mathrm{ex}} + \Delta\phi \to \Omega$. Hence, $\langle \Omega \rangle \geq 0$ (for Langevin systems) is Ref. [16]'s main result. From these connections, we see that our development not only provides new constraints on detailed fluctuations, but also extends these earlier results beyond Langevin systems.

Exposing these translations allows reformulating our detailed fluctuation theorems to apply to steady-state thermodynamics (SST). We have:

$$\begin{aligned} e^{\gamma(s^0|\boldsymbol{\mu}_{\mathrm{F}},x^0)-\gamma(s^{N-1}|\boldsymbol{\mu}_{\mathrm{R}},x^N)} \ e^{\Omega_{\mathrm{F}}+\Psi_{\mathrm{F}}} &= e^{\beta(Q_{\mathrm{hk}}-Q_{\mathrm{ex}})+\Delta S^{\mathrm{sys}}} \\ &= e^{\Delta S_{\mathrm{F}}^{\mathrm{tot}}} \ , \end{aligned}$$

where:

$$\Delta S^{\mathrm{sys}} \equiv -\ln \frac{\Pr(s^{N-1}|\boldsymbol{\mu}_{\mathrm{R}}, x^N)}{\Pr(s^0|\boldsymbol{\mu}_{\mathrm{F}}, x^0)} \ ,$$

when we choose $\langle \boldsymbol{\mu}_{\mathrm{R}}| = \langle \boldsymbol{\mu}_{\mathrm{F}}| \, \mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|x^0\mathbf{x})}$ and where $S_{\mathrm{F}}^{\mathrm{tot}}$ is the total change in entropy in forward time. This yields:

$$\frac{\Pr(\mathcal{S}_{0:N} = s^0\mathbf{s}|\mathcal{S}_{-1} \sim \boldsymbol{\mu}_{\mathrm{F}}, X_{0:N} = x^0\mathbf{x})}{\Pr(\mathcal{S}_{0:N} = s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}|\mathcal{S}_{-1} \sim \boldsymbol{\mu}_{\mathrm{R}}, X_{0:N} = x^N\mathbf{x}^{\mathrm{R}})} = e^{\Delta S_{\mathrm{F}}^{\mathrm{tot}}} \ .$$

And so, we immediately see that:

$$\langle e^{-\Delta S_{\mathrm{F}}^{\mathrm{tot}}} \rangle_{\Pr(\mathcal{S}_{0:N}=s^0\mathbf{s}|\mathcal{S}_{-1}\sim\boldsymbol{\mu}_{\mathrm{F}},X_{0:N}=x^0\mathbf{x})} = 1 \ .$$

This extends the validity of Ref. [53]'s general integral fluctuation theorem beyond Langevin dynamics. Since the total change in entropy is time asymmetric—$\Delta S_{\mathrm{F}}^{\mathrm{tot}} = -\Delta S_{\mathrm{R}}^{\mathrm{tot}}$—we obtain the most direct CFT generalization valid outside of detailed balance:

$$\frac{\Pr(\Delta S^{\mathrm{tot}}|\boldsymbol{\pi}_{\boldsymbol{x}^{\mathbf{0}}}, x^0\mathbf{x}x^N)}{\Pr(-\Delta S^{\mathrm{tot}}|\boldsymbol{\pi}_{\boldsymbol{x}^{\boldsymbol{N}}}, x^N\mathbf{x}^{\mathrm{R}}x^0)} = e^{\Delta S^{\mathrm{tot}}} \ . \tag{34}$$

Again, this does not invoke a dual, unphysical dynamic. Equation (34) has been reported previously in various settings; see, e.g., Eq. (21) of Ref. [23] and Eq. (43) of Ref. [51]. The result gives a detailed fluctuation relation for the change in total entropy production when transitioning between steady states.

The new detailed fluctuation theorem of Eq. (28) for joint distributions goes further in refining SST. If starting in a steady state and executing a protocol in an isothermal environment, we find that:

$$\frac{\Pr(W_{\mathrm{ex}}, Q_{\mathrm{hk}}|\boldsymbol{\pi}_{\boldsymbol{x}^{\mathbf{0}}}, x^0\mathbf{x}x^N)}{\Pr(-W_{\mathrm{ex}}, -Q_{\mathrm{hk}}|\boldsymbol{\pi}_{\boldsymbol{x}^{\boldsymbol{N}}}, x^N\mathbf{x}^{\mathrm{R}}x^0)} = e^{\beta Q_{\mathrm{hk}}} e^{\beta W_{\mathrm{ex}}} \ .$$

This novel relation gives strong constraints on the thermodynamic behavior of systems driven between NESSs, since it constrains the *joint* distribution for excess work and housekeeping heat. Moreover, nonsteady-state additions to free energy are predicted when an experiment does not start in steady state.

In the special case of time-symmetric driving—$x^0\mathbf{x}x^0 = x^0\mathbf{x}^{\mathrm{R}}x^0 = x^0x^1x^2\ldots x^2x^1x^0$—and starting from an equilibrium distribution, the preceding expression reduces to a useful comparison between excess work values achieved by the single time-symmetric protocol:

$$\frac{\Pr(W_{\mathrm{ex}}, Q_{\mathrm{hk}})}{\Pr(-W_{\mathrm{ex}}, -Q_{\mathrm{hk}})} = e^{\beta Q_{\mathrm{hk}}} e^{\beta W_{\mathrm{ex}}} \ .$$

Similar results were recently derived in Ref. [54] under more restrictive assumptions for underdamped Langevin systems.

This all said, one must use caution and not always identify Ψ with βQ_{hk}. Most importantly, not all sources of irreversibility are naturally characterized as "heat". Thinking of irreversibility on its own dynamical terms is best.

F. Integral fluctuation theorems

Integral fluctuation theorems in the absence of detailed balance, starting arbitrarily far from equilibrium, also follow straightforwardly. One generalization of the integral fluctuation theorem [55] is:

$$\begin{aligned}&\left\langle e^{-\beta W_{\text{diss}}-\Psi}\right\rangle_{\Pr(s_{0:N}|\boldsymbol{\mu}_{\text{F}},\mathbf{x})} \\ &\quad= \sum_{s_{0:N}\in\boldsymbol{\mathcal{S}}^N} \Pr(s_{0:N}|\boldsymbol{\mu}_{\text{F}},\mathbf{x}) e^{-\beta W_{\text{diss}}-\Psi} \\ &\quad= \sum_{s_{0:N}\in\boldsymbol{\mathcal{S}}^N} \Pr(s^0 \xleftarrow{x^1} \cdots \xleftarrow{x^N} \boldsymbol{\mu}(\boldsymbol{\mu}_{\text{F}},\mathbf{x}) | \boldsymbol{\mu}(\boldsymbol{\mu}_{\text{F}},\mathbf{x}), \mathbf{x}^{\text{R}}) \\ &\quad= 1 \ . \end{aligned} \tag{35}$$

If the input is stochastic, then averaging over the input also gives:

$$\left\langle e^{-\beta W_{\text{diss}}-\Psi}\right\rangle_{\Pr(x_{0:N},s_{0:N}|\boldsymbol{\mu}_{\text{F}})} = 1 \ .$$

Note that this relation does *not* require the system to be in steady state at any time.

From the concavity of the exponential function, it is tempting to assert a corresponding generalized Second Law of SST as:

$$\langle W_{\text{diss}} \rangle \geq - \langle Q_{\text{hk}} \rangle \ . \tag{36}$$

Although Eq. (36) is true, notably it is neither a strong nor useful bound. Let's address this. Note that:

$$\left\langle e^{-\Psi}\right\rangle_{\Pr(s_{0:N}|\boldsymbol{\mu}_{\text{F}},\mathbf{x})} = 1 \tag{37}$$

and:

$$\left\langle e^{-\beta W_{\text{diss}}}\right\rangle_{\Pr(s_{0:N}|\boldsymbol{\mu}_{\text{F}},\mathbf{x})} = 1 \ . \tag{38}$$

Both follow from the normalization of probabilities of the conjugate dynamic. Therefore, $\langle \Psi \rangle \geq 0$ *and* $\langle W_{\text{diss}} \rangle \geq 0$. And, hence $\langle Q_{\text{hk}} \rangle \geq 0$ as shown in Ref. [56]. So, Eq. (36) is devoid of utility. Nevertheless, Eq. (35) puts a novel constraint on the joint distributions of W_{diss} and Ψ.

Introducing an artificial conjugate dynamic following Ref. [50] and following the derivation there with ϕ/β in place of E, when starting in the steady state distribution $\boldsymbol{\pi}_{\boldsymbol{x_0}}$, we can show that:

$$\left\langle e^{-\Omega}\right\rangle_{\Pr(s_{0:N}|\boldsymbol{\pi}_{\boldsymbol{x_0}},\mathbf{x})} = 1 \ , \tag{39}$$

which implies the restriction $\langle \Omega \rangle \geq 0$. Despite similar appearance, this result has meaning beyond Crooks' derivation of the Jarzynski equality, as it now also applies *atop nonequilibrium steady states*. Recall that Ω has general meaning as in Eq. (7): $\Omega = \beta W_{\text{ex}} = -\beta Q_{\text{ex}} + \Delta\phi$. So, Eq. (39) becomes:

$$\left\langle e^{-\beta W_{\text{ex}}}\right\rangle_{\Pr(s_{0:N}|\boldsymbol{\pi}_{\boldsymbol{x_0}},\mathbf{x})} = 1 \ . \tag{40}$$

Effectively, this is Ref. [16]'s relation that, with our sign convention for Q_{ex}, implies:

$$\begin{aligned}\langle \Omega \rangle &= \langle -\beta Q_{\text{ex}} + \Delta\phi \rangle \\ &\geq 0 \ , \end{aligned} \tag{41}$$

for processes that start in steady state.

However, using Eq. (38), we find a more precise constraint on expected excess entropy production whether or not the system starts in steady state:

$$\langle \Omega \rangle \geq \Delta \langle \gamma \rangle \ , \tag{42}$$

where the RHS can be positive or negative, but can only be negative if the system starts out of steady state. When starting in a steady state, this yields:

$$\langle \Omega \rangle \geq \langle \gamma_{\text{final}} \rangle \tag{43}$$

$$= D_{\text{KL}}\left[\Pr(\mathcal{S}_t|\mathcal{S}_0 \sim \boldsymbol{\pi}_{\boldsymbol{x_0}}, x_{1:t+1}) \,||\, \boldsymbol{\pi}_{\boldsymbol{x_t}}\right] \ , \tag{44}$$

which is a stronger constraint than the previous result of Eq. (41), since the RHS is always positive for $\Pr(\mathcal{S}_t|\mathcal{S}_0 \sim \boldsymbol{\pi}_{\boldsymbol{x_0}}, x_{1:t+1}) \neq \boldsymbol{\pi}_{\boldsymbol{x_t}}$. Equation (44) extends the validity of the main result obtained in Ref. [57] to now include the possibility of starting in a nonequilibrium steady state and allowing for nondetailed-balanced dynamics. (Note that 'W_{diss}' in Ref. [57] corresponds to our W_{ex}—it is excess work that is not necessarily yet dissipated.)

Integral fluctuation theorems for systems with controlled or intrinsic feedback also directly follow, as we now show, extending the theory of feedback control to the setting of transitions between NESSs.

G. Fluctuation theorems with an auxiliary variable

Actions made by a complex thermodynamic system can couple back from the environment to influence the system's future input. To achieve this, the system may express an auxiliary random variable Y_t—the current "output" that takes on the values $y \in \mathcal{Y}$ and is instantaneously energetically mute, but may influence the future input and so does have energetic relevance.

The variable Y_t could be measurement, output, or any other auxiliary variable that influences the state or input sequences. To be concise, we introduce a shorthand for the time-ordered sequences of random variables: $\overrightarrow{X} \equiv X_{0:N}$, $\overrightarrow{S} \equiv \mathcal{S}_{0:N}$, and $\overrightarrow{Y} \equiv Y_{0:N}$. And, for particular realizations of the sequences: $\overrightarrow{x} \equiv x_{0:N}$, $\overrightarrow{s} \equiv s_{0:N}$, and $\overrightarrow{y} \equiv y_{0:N}$. When time reversing realizations, we let $\overleftarrow{x} = x^{N-1}x^{N-2}\cdots x^1 x^0$ and $\overleftarrow{s} = s^{N-1}s^{N-2}\cdots s^1 s^0$. To clarify further, $\overrightarrow{x}$ appearing inside a probability implies $\overrightarrow{X} = \overrightarrow{x}$ and $\overleftarrow{s}$ appearing inside a probability implies $\overrightarrow{S} = \overleftarrow{s}$.

We quantify how much the auxiliary variable is independently informed from the state sequence—beyond what could be known if given only the initial distribution over states and the driving history—via the unaveraged conditional mutual information:

$$\begin{aligned} i[\overrightarrow{s};\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}}] &\equiv \ln \frac{\Pr(\overrightarrow{s},\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}})\Pr(\overrightarrow{s}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}})} \\ &= \ln \frac{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_{\mathrm{F}})\Pr(\overrightarrow{s}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}})} ~. \end{aligned}$$

Note that averaging over the input, state, and auxiliary sequences gives the familiar conditional mutual information: $\mathrm{I}[\overrightarrow{S};\overrightarrow{Y}|\overrightarrow{X},\boldsymbol{\mu}_{\mathrm{F}}] = \langle i[\overrightarrow{s};\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}}]\rangle_{\Pr(\overrightarrow{x},\overrightarrow{s},\overrightarrow{y}|\boldsymbol{\mu}_{\mathrm{F}})}$.

As detailed in App. B:

$$e^{\beta W_{\mathrm{diss}} + i[\overrightarrow{s};\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}}] + \Psi} = \frac{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_{\mathrm{F}})\Pr(\overleftarrow{s}|\overleftarrow{x},\boldsymbol{\mu}_{\mathrm{R}})} ~,$$

where $\boldsymbol{\mu}_{\mathrm{R}} = \boldsymbol{\mu}(\boldsymbol{\mu}_{\mathrm{F}},\overrightarrow{x})$. This leads directly to the integral fluctuation theorem:

$$\left\langle e^{-\beta W_{\mathrm{diss}} - i[\overrightarrow{s};\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}}] - \Psi} \right\rangle_{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_{\mathrm{F}})} = 1 ~. \qquad (45)$$

However, as before, the resulting bound on $\langle W_{\mathrm{diss}}\rangle$ is not the tightest possible. Alternatively, we can invoke the normalization of conjugate dynamics to show:

$$\left\langle e^{-\beta W_{\mathrm{diss}} - i[\overrightarrow{s};\overrightarrow{y}|\overrightarrow{x},\boldsymbol{\mu}_{\mathrm{F}}]} \right\rangle_{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_{\mathrm{F}})} = 1 ~. \qquad (46)$$

This implies a new lower bound for the revised Second Law of Thermodynamics:

$$\langle W_{\mathrm{diss}}\rangle \geq -k_{\mathrm{B}}T ~ \mathrm{I}[\overrightarrow{S};\overrightarrow{Y}|\overrightarrow{X},\boldsymbol{\mu}_{\mathrm{F}}] ~, \qquad (47)$$

enabled by the conditional mutual information between state-sequence and auxiliary sequence, given input-sequence. Notably, this relation holds arbitrarily far from equilibrium and allows for the starting and ending distributions to be nonsteady-state.

We may also be interested in the unaveraged unconditioned mutual information between the auxiliary variable sequence and the joint input–state sequence. Then, using:

$$i[\overrightarrow{y};\overrightarrow{xs}|\boldsymbol{\mu}_{\mathrm{F}}] \equiv \ln \frac{\Pr(\overrightarrow{x},\overrightarrow{s},\overrightarrow{y}|\boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y}|\boldsymbol{\mu}_{\mathrm{F}})\Pr(\overrightarrow{x},\overrightarrow{s}|\boldsymbol{\mu}_{\mathrm{F}})} ~,$$

we find that, in general:

$$\langle W_{\mathrm{diss}}\rangle \geq -k_{\mathrm{B}}T ~ \mathrm{I}[\overrightarrow{Y};\overrightarrow{XS}|\boldsymbol{\mu}_{\mathrm{F}}] \qquad (48)$$

and when starting in steady-state:

$$\langle\Omega\rangle \geq -\mathrm{I}[\overrightarrow{Y};\overrightarrow{XS}|\boldsymbol{\pi}_{x^0}] ~. \qquad (49)$$

One can now continue in this fashion to successively derive a seeming unending sequence of fluctuation theorems. Let's stop, though, with one more and discuss its interpretations and applications.

As a final set of example integral fluctuation theorems, we follow Ref. [3] in defining:

$$i_{\mathrm{SU}} \equiv \ln \frac{\Pr(\overrightarrow{y},\overrightarrow{s}|\boldsymbol{\mu}_0)}{\Pr(\overrightarrow{y}|\boldsymbol{\mu}_0)\Pr(\overrightarrow{s}|\boldsymbol{\mu}_0,\overrightarrow{x})} ~.$$

(This is Ref. [3]'s I_{C}, if $\boldsymbol{\mu}_0 \to \boldsymbol{\pi}_{x^0}$.) Technically speaking, this is not a mutual information, even upon averaging. Then, we arrive at the integral fluctuation theorems:

$$\left\langle e^{-W_{\mathrm{diss}} - i_{\mathrm{SU}} - \Psi} \right\rangle_{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_0)} = 1$$

and

$$\left\langle e^{-W_{\mathrm{diss}} - i_{\mathrm{SU}}} \right\rangle_{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\mu}_0)} = 1 ~.$$

When starting from a steady-state distribution, we have the most direct generalization of Ref. [3]'s feedback control result, but extended to not require detailed balance:

$$\left\langle e^{-\Omega - i_{\mathrm{SU}}} \right\rangle_{\Pr(\overrightarrow{s},\overrightarrow{y},\overrightarrow{x}|\boldsymbol{\pi}_{\boldsymbol{x}^{\mathbf{0}}})} = 1 ~.$$

When starting from a NESS, this suggests that:

$$\langle\Omega\rangle \geq -\mathrm{I}_{\mathrm{SU}} ~. \qquad (50)$$

When the dynamics have detailed balance, this naturally reduces to the well known results of Ref. [3] and others: $\langle W\rangle \geq \Delta F_{\mathrm{eq}} - k_{\mathrm{B}}T ~ \mathrm{I}_{\mathrm{SU}}$.

In the feedback control setting, Y_n is said to be the random variable for measurements at time n. This suggests that Y_n is a function of $\mathcal{S}_n$ and the outcome of Y_n effectively induces different Markov chains over the states since X_{n+1} is a function of Y_n—i.e., $x_{n+1}(y_n(s_n))$.

With our interest in complex autonomous systems, we note that our results give new bounds on the Second Law of Thermodynamics for highly structured complex systems strongly coupled to an environment. A preliminary application of this was presented in Ref. [39]. We offer our own in the next section. Analysis of thermodynamic systems with the agency to influence their environmental input, via some kind of coupling or feedback, say, will likely benefit from our extended theory.

How can we reconcile this with other inequalities without auxiliary Y? The other inequalities used averages of variable occurrence already *conditioned* on $\vec{x}$. However, if input x and states s can influence each other dynamically through auxiliary y, then averaging over their joint dynamic allows less dissipation than the traditional Second Law suggests.

If $\mathcal{S}$ represents the random variable for one subset of a system's degrees of freedom, and Y represents the random variable for another subset of a system's degrees of freedom, then the intrinsic nonextensivity of the thermodynamic entropy:

$$S(\mathcal{S}, Y|X) = S(\mathcal{S}|X) + S(Y|X) - k_\text{B}\,\text{I}(\mathcal{S}; Y|X)$$

goes a long way towards explaining why the physics of information has stimulated a recent resurgence of Maxwellian demonology. This viewpoint is taken up in Ref. [12, Ch. 9] and will be further developed elsewhere.

VI. PARTIALLY CONTROLLED DRIVING BETWEEN DESIRED SETS OF CONFIGURATIONS

There are countless engineering and biological situations where we wish to induce a metastable macrostate of an active energy-consuming system. For example, in an active CMOS circuit, we may desire the output of our computation to be readable as a '1' rather than a '0'. However, we may not have full control of the system. In a biological system, to take another example, we may try to influence an assembly of multi-cellular structures, even though many aspects of cellular upkeep are beyond our control.

Let us now consider the work that will be dissipated in driving a system between two (possibly overlapping) sets of states, $\mathbf{I} \subset \boldsymbol{\mathcal{S}}$ and $\mathbf{II} \subset \boldsymbol{\mathcal{S}}$.

Notice that the emphasis is changing somewhat in this section compared to the previous sections: Here, we are interested in the work dissipated driving between two sets of configurations rather than driving between steady states. In fact, the driving may start and end arbitrarily far from even any steady state. Nevertheless, the concepts of *excess* and *housekeeping* continue to play an important practical role, to distinguish the thermodynamic roles played by those forces that we have control over versus those that we do not, respectively. Some forces out of our control keep the system out of equilibrium—these generate the housekeeping energy fluxes. However, the energy dissipated due to those forces that we exert and have control over can still be tracked even arbitrarily far from steady state, via the excess and dissipated work. This is clearly of interest if we want to minimize our exertion while successfully influencing a complex system.

The scenario of influencing a complex system to transition from one subset of functionally relevant configurations to another is of vast practical importance. The following development of this somewhat parallels that addressed to self-replication [27–29]. However, here we allow the possibility that only some of the driving influences are controllable. This generalization is important, for example, in (i) understanding the limitations of how complex systems can influence each other given limited modes of influence and uncontrollable homeostatic upkeep and (ii) for identifying the energy each must expend on that specific control task. We find that when uncontrollable forces produce nondetailed balanced dynamics, the built-in irreversibility can be leveraged to influence the desired transitions while exerting less energy for control. The self-replication results re-emerge as a special case of detailed-balanced dynamics, where $\Psi = 0$ and there is no need to distinguish between controllable and uncontrollable forces.

To proceed, we must quite explicit, and so we start with several definitions. We are interested in the work dissipated when driving a system between two (possibly overlapping) sets of states, $\mathbf{I} \subset \boldsymbol{\mathcal{S}}$ and $\mathbf{II} \subset \boldsymbol{\mathcal{S}}$. To calculate the dissipated work in these cases, we need to condition on the state starting in a type $\mathbf{I}$ configuration and also ending in a type $\mathbf{II}$ configuration. We also need to consider the probability distributions induced over $\mathbf{I}$ and $\mathbf{II}$ when the system is initiated (via whatever preparation) arbitrarily far from equilibrium and found to be in $\mathbf{I}$ and subsequently when the system has evolved from type $\mathbf{I}$ into type $\mathbf{II}$ via the driving protocol $\mathbf{x}$. Recall that $\boldsymbol{\mu}_\text{F}$ can be an arbitrary initial distribution over the system's state set $\boldsymbol{\mathcal{S}}$. Let $[\boldsymbol{\mu}_\text{F}]_\mathbf{I}$ be the restriction of the start distribution to the subset of states $\mathbf{I}$, properly renormalized in probability; i.e.:

$$[\boldsymbol{\mu}_\text{F}]_\mathbf{I} = \frac{\sum_{s \in \mathbf{I}} \langle \boldsymbol{\mu}_\text{F} | s \rangle \, \langle s |}{\sum_{s \in \mathbf{I}} \langle \boldsymbol{\mu}_\text{F} | s \rangle} ~,$$

where $\langle s|$ is the delta-distributed probability distribution over states with full weight at s. Similarly, $[\boldsymbol{\mu}([\boldsymbol{\mu}_\text{F}]_\mathbf{I}, \mathbf{x})]_\mathbf{II}$ is the probability distribution over $\mathbf{II}$ induced by having started in a type $\mathbf{I}$ state and having ended in a type $\mathbf{II}$ state via the driving protocol $\mathbf{x}$; i.e.:

$$[\boldsymbol{\mu}([\boldsymbol{\mu}_\text{F}]_\mathbf{I}, \mathbf{x})]_\mathbf{II} = \frac{\sum_{s \in \mathbf{II}} \langle [\boldsymbol{\mu}_\text{F}]_\mathbf{I} | \mathsf{T}^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}} | \mathbf{x})} | s \rangle \, \langle s |}{\sum_{s \in \mathbf{II}} \langle [\boldsymbol{\mu}_\text{F}]_\mathbf{I} | \mathsf{T}^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}} | \mathbf{x})} | s \rangle} ~.$$

When the analytics are intractable, these distributions can be experimentally inferred.

The actual work dissipated in any such driven transition varies due to uncertainty in the start and end states and, also, due to the stochastic system dynamics. Nevertheless, the distribution of dissipated-work values turns out to be constrained by several intuitive factors, as we now show. Calculating, using the result of Eq. (27), we find:

$$
\begin{aligned}
&\left\langle e^{-\beta W_{\mathrm{diss}}-\Psi}\right\rangle_{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{0:N}=s^{0:N}|\mathcal{S}_0\in\mathbf{I},\,\mathcal{S}_{N-1}\in\mathbf{II},\,\mathbf{x})}\\
&\quad=\sum_{s^{0:N}\in\boldsymbol{\mathcal{S}}^N}\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{0:N}=s^{0:N}|\mathcal{S}_0\in\mathbf{I},\,\mathcal{S}_{N-1}\in\mathbf{II},\,\mathbf{x})\,\frac{\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_{N:2N}=s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}s^0|\mathbf{x}^{\mathrm{R}})}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{0:N}=s^{0:N}|\mathbf{x})}\\
&\quad=\sum_{s^{0:N}\in\boldsymbol{\mathcal{S}}^N}\frac{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{0:N}=s^{0:N}|\mathbf{x})\,\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_0\in\mathbf{I},\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_{0:N}=s^{0:N},\mathbf{x})}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_0\in\mathbf{I}|\mathbf{x})\,\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_0\in\mathbf{I},\mathbf{x})}\,\frac{\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_{N:2N}=s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}s^0|\mathbf{x}^{\mathrm{R}})}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{0:N}=s^{0:N}|\mathbf{x})}\\
&\quad=\frac{1}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_0\in\mathbf{I},\mathbf{x})}\sum_{\substack{s^0\in\mathbf{I}\\ s^{N-1}\in\mathbf{II}\\ s^{1:N-1}\in\boldsymbol{\mathcal{S}}^{N-2}}}\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_{N:2N}=s^{N-1}\mathbf{s}^{\mathrm{R}}_{\leftarrow}s^0|\mathbf{x}^{\mathrm{R}})\\
&\quad=\frac{1}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_0\in\mathbf{I},\mathbf{x})}\sum_{\substack{s^0\in\mathbf{I}\\ s^{N-1}\in\mathbf{II}}}\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_N=s^{N-1},\,\mathcal{S}_{2N-1}=s^0|\mathbf{x}^{\mathrm{R}})\\
&\quad=\frac{\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_N\in\mathbf{II},\,\mathcal{S}_{2N-1}\in\mathbf{I}|\mathbf{x}^{\mathrm{R}})}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_0\in\mathbf{I},\mathbf{x})}\\
&\quad=\frac{\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_{2N-1}\in\mathbf{I}|\mathcal{S}_N\in\mathbf{II},\mathbf{x}^{\mathrm{R}})}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_0\in\mathbf{I},\mathbf{x})}\ .
\end{aligned}
\tag{51}
$$

We used the facts that $\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_0\in\mathbf{I},\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_{0:N}=s^{0:N},\mathbf{x})=\delta_{s^0\in\mathbf{I}}\,\delta_{s^{N-1}\in\mathbf{II}}$, that $\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_0\in\mathbf{I}|\mathbf{x})=1$ and, similarly, that $\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_N\in\mathbf{II}|\mathbf{x}^{\mathrm{R}})=1$.

The ratio of probabilities on the RHS of Eq. (51) has a rather straightforward interpretation: The denominator is the probability of transitioning to a configuration of type **II** via the control protocol $\mathbf{x}$, when starting from a configuration of type **I**, given that the initial preparation had induced the measure $[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}$ over **I**. For shorthand, let us call this: $\Pr(\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})$. The numerator is the probability of transitioning back to a type **I** configuration via the reverse control protocol $\mathbf{x}^{\mathrm{R}}$, given a successful $\mathbf{x}$-driven transition from **I** to **II**. For shorthand, let us call this: $\Pr(\mathbf{II}\xrightarrow{\mathbf{x}^{\mathrm{R}}}\mathbf{I}\,|\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})$. Using this shorthand, Eq. (51) can be rewritten as:

$$
\left\langle e^{-\beta W_{\mathrm{diss}}-\Psi}\right\rangle_{\Pr(\overrightarrow{s}|\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}=\frac{\Pr(\mathbf{II}\xrightarrow{\mathbf{x}^{\mathrm{R}}}\mathbf{I}\,|\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}{\Pr(\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}\ . \tag{52}
$$

Since $\langle Y\rangle\geq-\ln\langle e^{-Y}\rangle$ for any real-valued random-variable Y [58], Eq. (51) leads us to:

$$
\langle\beta W_{\mathrm{diss}}+\Psi\rangle_{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{0:N}=s^{0:N}|\mathcal{S}_0\in\mathbf{I},\,\mathcal{S}_{N-1}\in\mathbf{II},\,\mathbf{x})}\geq-\ln\frac{\Pr_{\mathcal{S}_N\sim[\boldsymbol{\mu}([\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}},\mathbf{x})]_{\mathbf{II}}}(\mathcal{S}_{2N-1}\in\mathbf{I}|\mathcal{S}_N\in\mathbf{II},\mathbf{x}^{\mathrm{R}})}{\Pr_{\mathcal{S}_0\sim[\boldsymbol{\mu}_{\mathrm{F}}]_{\mathbf{I}}}(\mathcal{S}_{N-1}\in\mathbf{II}|\mathcal{S}_0\in\mathbf{I},\mathbf{x})}\ . \tag{53}
$$

The result can be written more intuitively as:

$$
\langle W_{\mathrm{diss}}\rangle_{\Pr(\overrightarrow{s}|\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}\geq\beta^{-1}\ln\frac{\Pr(\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}{\Pr(\mathbf{II}\xrightarrow{\mathbf{x}^{\mathrm{R}}}\mathbf{I}\,|\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}-\beta^{-1}\,\langle\Psi\rangle_{\Pr(\overrightarrow{s}|\mathbf{I}\xrightarrow{\mathbf{x}}\mathbf{II})}\ . \tag{54}
$$

This leads directly to several useful conclusions. First, this result implies that more work will have been dissipated if it is more probable for the control protocol $\mathbf{x}$ to take **I** to **II** than for the reverse protocol $\mathbf{x}^{\mathrm{R}}$ to take **II** back to **I**. Roughly speaking, more hysteresis implies that more work must have been dissipated.

However, the result further suggests an intriguing opportunity. With a cleverly designed control protocol, the housekeeping irreversibility built into the controllable dynamics can be leveraged to perform the desired task, enabling less of the energy input to the controller to be dissipated than would be required if the dynamic were detailed balance. In other words, via protocol $\mathbf{x}$ the controller may seek to initiate a situation in which the built-in homeostatic forces irreversibly bring the system towards the desired configuration at the cost of greater housekeeping entropy production. Crucially, *this greater housekeeping entropy production can offset the dissipated energy that the controller would otherwise need to exert.* Although space precludes us from fleshing such consequences out, in mentioning at least one we hope to indicate the kind of application of these results that we foresee in the near future.

Thus, it is clear that our result is useful in the context of control and leveraging the natural dynamics of complex systems. Let us be concrete, though, about how it compares with recent results on the thermodynamics of replication in Refs. [27–29].

There are two obvious ways to get from our Eq. (54) to the thermodynamic core behind the replication theory. One approach is to assume that the controllable dynamics have detailed balance. (This is equivalent to assuming that there are no uncontrolled sources that power the system.) Then $\Psi = 0$, and the dissipated work is bounded simply by the first term on the RHS of Eq. (54). Using Eq. (19) together with $\gamma = \phi - h$, and the fact that $W_{\text{ex}} = W - \Delta F_{\text{eq}}$ and $\beta^{-1}\phi = E - F_{\text{eq}}$, when the fundamental system dynamics have detailed balance, we find that:

$$\langle W - \Delta E + \beta^{-1}\Delta h \rangle_{\Pr(\overrightarrow{s}|\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})} \overset{?}{\geq} \beta^{-1} \ln \frac{\Pr(\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})}{\Pr(\mathbf{II} \xrightarrow{\mathbf{x}^{\text{R}}} \mathbf{I}\,|\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})}$$

is true—but only when all driving of a detailed-balanced system is due to work performed by the controller, as in Eq. (3) of Ref. [28].

Alternatively, if we do not care about isolating the work that must be dissipated by a controller acting on an actively powered complex system, but only care about the overall entropy production in the universe, then there is no need to distinguish between controllable and uncontrollable forces. In such a case, we lump the housekeeping and excess heat together. We then use Eq. (20) and $Q = Q_{\text{ex}} - Q_{\text{hk}}$ (given our sign conventions for Q, Q_{hk}, and Q_{ex}), yielding $W_{\text{diss}} + \beta^{-1}\Psi = -Q_{\text{ex}} + \beta^{-1}\Delta h + Q_{\text{hk}} = -Q + \beta^{-1}\Delta h$, to find the net entropy-production inequality:

$$\langle -\beta Q + \Delta h \rangle_{\Pr(\overrightarrow{s}|\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})} + \ln \frac{\Pr(\mathbf{II} \xrightarrow{\mathbf{x}^{\text{R}}} \mathbf{I}\,|\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})}{\Pr(\mathbf{I} \xrightarrow{\mathbf{x}} \mathbf{II})} \geq 0 \ . \tag{55}$$

This again resembles Eq. (8) of Ref. [27], although our derivation shows that Eq. (55) (with Q containing both excess and housekeeping heat) is valid even when the system dynamics are nondetailed-balanced.

Given these similarities, let us reiterate why our result in Eq. (54) is an important extension of the original theoretical basis for self-replication. To emphasize, our result is useful when the observed controllable dynamic is nondetailed-balanced due to uncontrolled forces that constantly drive the system of interest away from equilibrium. Indeed, our result provocatively suggests that less energy exertion is required of the controller, if it leverages the built-in irreversibilities in the system dynamic.

Note that, unlike the integral fluctuation theorems that average over the full state-space, we cannot invoke the artificial dual dynamics for the analogous result without Ψ, unless we accept the unphysical dual dynamics' probabilities in the final ratio of probabilities. The situation is similar to that of the CFT in the face of nondetailed-balanced dynamics: the housekeeping irreversibility Ψ must be accounted for if we want to discuss physically meaningful probabilities. Here, at least, the entrance of the housekeeping irreversibility Ψ appears as an opportunity. A controller can induce a desired transition using built-in irreversibilities to offset energy dissipation that the controller itself would otherwise need to exert.

VII. NESS TRANSITIONS IN NEURONAL DYNAMICS

Our nervous systems present a rich testing ground for practical theories of nonequilibrium thermodynamics. In them one finds no shortage of homeostasis (biologically maintained NESSs) nor of adaptation (switching between NESSs) at multiple timescales, corresponding to different types of memory and learning [59, 60]. Moreover, our brains perform their information processing tasks with remarkable energy efficiency [61, 62].

Neurons actively maintain a nonequilibrium state to readily transmit and process information. ATP is consumed by ion pumps that maintain a relatively constant set of ionic concentration gradients across the neuronal cell membrane [63]. The work of the ion pumps is ongoing and contributes significantly to the housekeeping energy consumption of the neural membrane [64].

While ion pumps and passive channels support ionic concentration gradient homeostasis, transmembrane voltage-gated ion channels leverage these nonequilibrium steady states to propel speedy information transmission along the cell membrane. Acting in concert, voltage-gated sodium ion channels and potassium ion channels form the active substrate that drives the evolution of membrane potentials in neurons [65]. Together, these voltage-gated channels are the primary generators of the action potentials or "spikes" that are believed to comprise the basic signals. Their collective patterns support information processing in neural circuits, cell assemblies, cell groups, and brain tissues [66].

In experiments, if the cell membrane is voltage clamped, then the voltage-gated channels approach a stationary distribution over their conformational states according to the effective energies of their biomolecular conformations at that voltage [67]. However, absent clamping, the channels influence their own voltage input dynamically through their current output. The result is spontaneous spiking patterns.

Thermodynamically, although the channels do not directly consume ATP (ion *pumps* do), the channels play an important role in the *dissipation* of the potential energy stored in the ionic concentration gradients. While the dissipation is necessary to transmit the information-bearing signals, the dynamics of the channels turns out to be nearly minimally wasteful for the task [61].

Although this is not the setting in which to analyze the full richness of ion channel dynamics and interactions, we use the sodium channel under different voltage-driving protocols to demonstrate relatively straightforwardly the insights on NESS transitions gained from our theoretical results. In particular, while potassium ion channels are somewhat structured, the sodium ion channel exhibits a more structured and so more illustrative dynamic over its coarse-grained state space of functional protein conformations.

Our goal here is not to analyze all of the thermodynamic costs of neural processing nor is it to illustrate the full array of theoretical results. Rather, we take on the more circumscribed task to demonstrate that the amount of work dissipated by an interesting biological system as it is steered to a new NESS can be analyzed, despite the many levels of complexity and dissipation acting in concert. This serves to emphasize why the excess work is relevant at all in the face of constant housekeeping dissipation that appears dominant. That is, the excess work and dissipated work are relevant since they specifically address a channel's thermodynamic efficiency when responding to environmental control. Concretely, how much of the exerted work was dissipated? The dissipated work W_{diss}. Moreover, all excess work is dissipated, if the system relaxes to its new steady state.

In short, *while the housekeeping heat Q_{hk} monitors the persistent cost of homeostasis, the excess work W_{ex} tracks the inefficiency of adaptive response to new stimuli.* Often, as the system processes and responds to its change of input, W_{ex} corresponds to a *cost of neural information processing*. Similarly, the excess environmental entropy production Ω summarizes the nonsystem entropy produced due to the channel's response to the control. This serves to illustrate that when systems are maintained out of equilibrium by an uncontrolled force (so that the controllable dynamic is irreversible), the excess work and environmental entropy produced in steering the system from one NESS to another jointly follow our extended CFT for irreversible processes.

A. Ion channel dynamics

For sodium ions to move through a dedicated Na$^+$ channel in the neural membrane, the channel's activation gates must be open and the deactivation gate must not yet plug the channel [63]. The rates of transitions among the conformational states have a highly nontrivial dependence on voltage across the cell membrane. Beyond this voltage dependence, while the activation gates act largely independently of one another, the inactivation gate cannot plug the channel until at least some of the activation gates are open. This contingent architecture was not yet captured by the relatively macroscopic differential equations introduced in the pioneering work of Hodgkin and Huxley [67]. Since then, however, it has been summarized by experimentally-motivated voltage-dependent Markov chains over the causally relevant conformational states [68]. Here, we follow the model implied in Ref. [63], whose voltage-dependent Markov chain we show in Fig. 1.

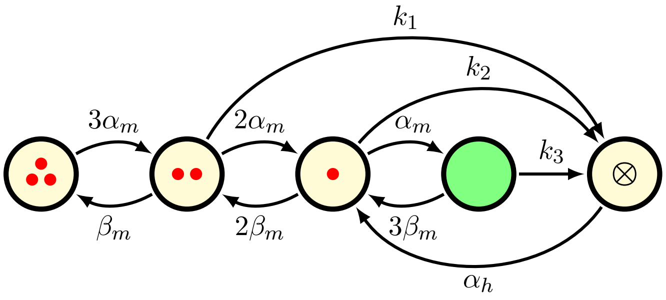


FIG. 1. Markov chain representation of the input-conditioned state-to-state rate matrix $G^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v)}$. Self-transitions are implied but, for tidiness, are not shown explicitly. These coarse-grainings of conformational states have biologically important functional interpretations. The number of (red) dots in each effective state corresponds to the number of activation gates that close off the channel. For example, when the channel is in the leftmost state, three activation gates are still active in blocking the channel. The solid (green) state corresponds to the channel being open. This is the only one of the five states in which sodium current can flow. The last state, marked $\otimes$, corresponds to channel inactivation by the inactivation gate—when the channel is plugged by its "ball and chain". Subsequent figures use the state numbering 1 through 5 for state identification, which corresponds to enumerating the states from left to right.

As the mesoscopic system of thermodynamic interest, the voltage-dependent Markov chain can be reinterpreted as a transducer that takes in voltage $v \in \mathbb{R}$ across the cell membrane as its input and makes transitions over its conformational state space $\boldsymbol{\mathcal{S}}$ according to an infinite set of transition matrices $\{\mathsf{T}^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v)}\}_{v\in\mathbb{R}}$.

Although the set is uncountable, the voltage-conditioned transition matrices are all described succinctly via time-independent functions of the voltage appearing in the transition elements; denoted α_m, β_m, and α_h. Time discretization of the continuous-time dynamic is straightforward and well behaved as $\Delta t \to 0$. If the voltage v is approximately constant during the infinitesimal interval Δt, then the state-to-state transition matrix is:

$$\begin{aligned} \mathsf{T}_{\Delta t}^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)} &= e^{(\Delta t) G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)}} \\ &\approx I + (\Delta t) G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)} ~, \end{aligned}$$

where I is the identity matrix and $G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)}$ is the infinitesimal generator of time evolution:

$$G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)} \equiv \begin{bmatrix} -3\alpha_m & 3\alpha_m & 0 & 0 & 0 \\ \beta_m & -(2\alpha_m + \beta_m + k_1) & 2\alpha_m & 0 & k_1 \\ 0 & 2\beta_m & -(\alpha_m + 2\beta_m + k_2) & \alpha_m & k_2 \\ 0 & 0 & 3\beta_m & -(3\beta_m + k_3) & k_3 \\ 0 & 0 & \alpha_h & 0 & -\alpha_h \end{bmatrix} ~. \tag{56}$$

Specifically, α_m, β_m, and α_h are voltage-dependent variables, as found in the Hodgkin and Huxley model [63, 67]:

$$\alpha_m(v) = \frac{(v + 40 \text{ mV})/10 \text{ mV}}{1 - \exp\left[-(v + 40 \text{ mV})/10 \text{ mV}\right]} ~,$$

$$\beta_m(v) = 4 \exp\left[-(v + 65 \text{ mV})/18 \text{ mV}\right] \qquad ,$$

and:

$$\alpha_h(v) = \tfrac{7}{100} \exp\left[-(v + 65 \text{ mV})/20 \text{ mV}\right] ~.$$

See Fig. 2. The reaction-rate constants are $k_1 = 6/25 \text{ ms}^{-1}$, $k_2 = 2/5 \text{ ms}^{-1}$, and $k_3 = 3/2 \text{ ms}^{-1}$.

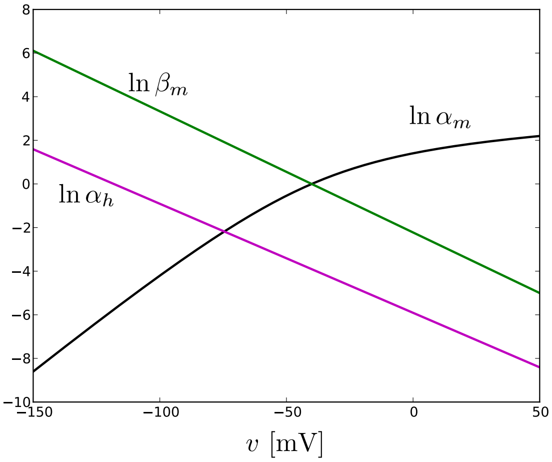


FIG. 2. Markov transition parameter voltage dependencies: $\ln \alpha_m$, $\ln \beta_m$, and $\ln \alpha_h$ versus cross-membrane voltage v. The plots show that at -100 mV, $\beta_m \gg \alpha_h \gg \alpha_m \approx 0$. At $+10$ mV, $\alpha_m \gg \beta_m \gg \alpha_h \approx 0$.

We see that $G^\dagger G \neq G G^\dagger$, and so the generator is not a normal matrix. Typically, this would preclude direct analysis. However, we can employ the new spectral decomposition methods from the meromorphic functional calculus [69, 70] to analytically calculate most, if not all, properties—e.g., dynamics, expected current, thermodynamics, information measures, and the like—about this model directly from the transition dynamic.

Since we are interested in thermodynamics, though, let us focus on determining the steady-state surprisals $\phi(x, s)$ of the conformational states s. For any persistent environmental input, the effective energies of the various conformational states are determined by their relative stationary occupation probability $\pi_x(s)$; according to Eq. (3), $\pi_x(s) = e^{-\phi(x,s)}$. The stationary distribution $\boldsymbol{\pi_v}$ induced by persistent v is the left eigenvector of $T_\tau^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)}$ associated with the eigenvalue of unity. Equivalently, and more convenient in this case, $\boldsymbol{\pi_v}$ is the left eigenvector of $G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)}$ associated with the eigenvalue of zero. Via $\langle \boldsymbol{\pi_v} | \, G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)} = \vec{0}$, we find that the steady state distribution for any persistent v is:

$$\begin{aligned} \boldsymbol{\pi_v} \propto \Bigg(& \frac{\beta_m}{3\alpha_m} ~,~ 1 ~,~ \frac{2\alpha_m + k_1}{2\beta_m} ~,~ \frac{\alpha_m}{2\beta_m} \left(\frac{2\alpha_m + k_1}{3\beta_m + k_3} \right) ~, \\ & \frac{1}{\alpha_h} \left[k_1 + \frac{2\alpha_m + k_1}{2\beta_m} \left(k_2 + \frac{k_3 \alpha_m}{3\beta_m + k_3} \right) \right] \Bigg) \\ \propto & ~ e^{-\boldsymbol{\phi}(v)} ~, \end{aligned}$$

which immediately yields the steady-state surprisals for conformational states at a constant environmental input v. The steady-state surprisal is shown for each conformational state in Fig. 3, as a function of the voltage-clamped membrane potential v.

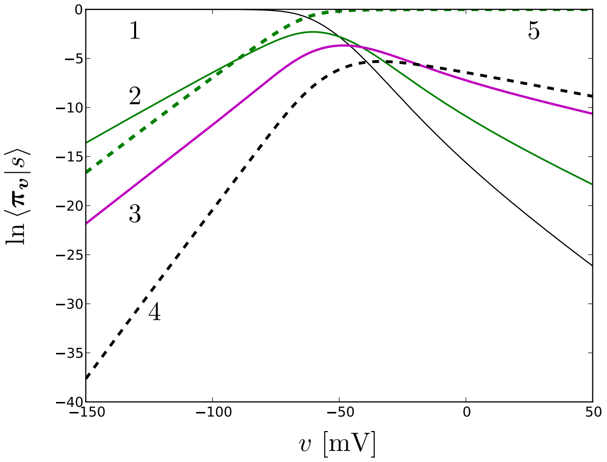


FIG. 3. Steady-state distribution voltage-dependence: negative steady-state surprisal $\ln \pi_v(s) = -\phi(v,s)$ for each conformational state. Each curve is labeled by the state-number to which it corresponds; recall Fig. 1. Note that −100 mV and +10 mV (relevant for later) are extremes in that $\boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}} \approx \boldsymbol{\delta_1}$ and $\boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{b}} \approx \boldsymbol{\delta_5}$.

B. Ion channel (ir)reversibility

Recall that detailed balance is the condition that, for states a and b and environmental input x:

$$\frac{\Pr(b \xrightarrow{x} a)}{\Pr(a \xrightarrow{x} b)} = \frac{\pi_x(a)}{\pi_x(b)} ~.$$

Interestingly, in this biologically inspired model, detailed balance is satisfied by *several, but not all* of the state-transition pairs.

For example, for very small Δt:

$$\begin{aligned}\frac{\Pr(1 \xrightarrow{v} 2)}{\Pr(2 \xrightarrow{v} 1)} &= \frac{3\alpha_m}{\beta_m} \\ &= \frac{\pi_v(2)}{\pi_v(1)} ~.\end{aligned}$$

That is, this transition pair satisfies detailed balance. Hence, all transitions between these states are completely reversible: $\Psi(2|1,v) = \Psi(1|2,v) = \Psi(22121112|2,v) = 0$.

However, this does not hold for other transition pairs. Consider transitions between states 2 and 3:

$$\begin{aligned}\frac{\Pr(2 \xrightarrow{v} 3)}{\Pr(3 \xrightarrow{v} 2)} &= \frac{\alpha_m}{\beta_m} \\ &\neq \frac{\pi_v(3)}{\pi_v(2)} \\ &= \frac{\alpha_m + k_1/2}{\beta_m} ~.\end{aligned}$$

Most other transitions also violate detailed balance.

Since detailed balance does not hold for the effective dynamic, the theory developed above is essential to analyzing the sodium ion channel thermodynamics. Moreover, the fact that $G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)}$ has null entries that are nonzero for its transpose implies that paths involving these transitions will be infinitely irreversible: $\Psi \to \infty$ for such paths as $\Delta t \to 0$; specifically, the transitions of $G^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|v)}$ with rates k_1 and k_3. Forbidden transitions are an extreme form of irreversibility that are nevertheless commonly observed for complex systems, as the ion channel so readily illustrates. In it, the asymmetry in allowed transitions can be traced to different *mechanisms* facilitating different paths through the state space.

Likely, the model's implied infinite irreversibility as $\Delta t \to 0$ is an artifact of experimental limitations in resolving a separation of timescales for these different mechanisms. However, insofar as the model is experimentally predictive for some large enough Δt, we should expect at least some of the housekeeping entropy production to be real. (More to the point, there is no reason to expect this system to obey detailed balance since the ion pumps actively maintain nonequilibrium conditions.) More careful experimental effort should be done to bound the actual housekeeping entropy production in these ion channels.

Fortunately, whether the irreversibility is truly infinite or just practically infinite does not matter much for the excess thermodynamics, although it will of course affect the calculated distribution of Ψ. And, therein lies one of the practical lessons of the present exercise: the thermodynamic costs of *controlling* a complex system can be predicted without needing to know all of the details of how it is maintained. All we really need to know for this purpose is the effective dynamics at the level we have access to control.

Profitably, this suggests that we do not need to limit ourselves to simple prototype models, but can apply the SST framework to gain real-world insights about control of quite sophisticated systems. We hope that our results stimulate investigation along these lines. By way of illustration, the following gives a simple example of controlling an ensemble of ion channels via enforced voltage across the cell membrane. The dissipated work there is thus the work dissipated by these channels in response to this particular driving force.

C. Step-function drive

With this understanding of ion channel NESSs, let's now turn to the thermodynamics induced by driving between them. We first consider the particular voltage protocol of $v_\mathrm{a} \equiv -100$ mV for all time except a $v_\mathrm{b} \equiv 10$ mV pulse for 5 ms starting at $t = 0$. This is an example of continuous-time dynamics and deterministic driving. The system begins in the steady state induced by the static environmental drive $v_\mathrm{a} = -100$ mV. The initial distribution over $\boldsymbol{\mathcal{S}}$ is thus $\boldsymbol{\mu}_0 = \boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}}$, where $\boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}}$ is the left eigenvector of $G^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v_\mathrm{a})}$ associated with the eigenvalue of zero.

During an epoch of fixed $v = V$, the net transition dynamic after τ ms becomes:

$$T_\tau^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v=V)} = e^{\tau G^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v=V)}} \ .$$

Therefore, the distribution over states induced by the driving protocol is:

$$\langle \boldsymbol{\mu}_t | = \begin{cases} \langle \boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}} | & \text{for } t \leq 0 \\ \langle \boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}} | \, e^{tG_\mathrm{b}} & \text{for } 0 < t \leq 5 \text{ ms} \\ \langle \boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}} | \, e^{5G_\mathrm{b}} e^{(t-5)G_\mathrm{a}} & \text{for } t > 5 \text{ ms} \end{cases} \ , \qquad (57)$$

where, for brevity, we defined: $G_\mathrm{a} \equiv G^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v_\mathrm{a})}$ and $G_\mathrm{b} \equiv G^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v_\mathrm{b})}$. These are especially useful when expressing the rate matrix via its spectral decomposition, using the methods of Refs. [69, 70]. Besides the zero eigenvalue, there are only four other eigenvalues of G that are determined via $\det(\lambda I - G) = 0$.

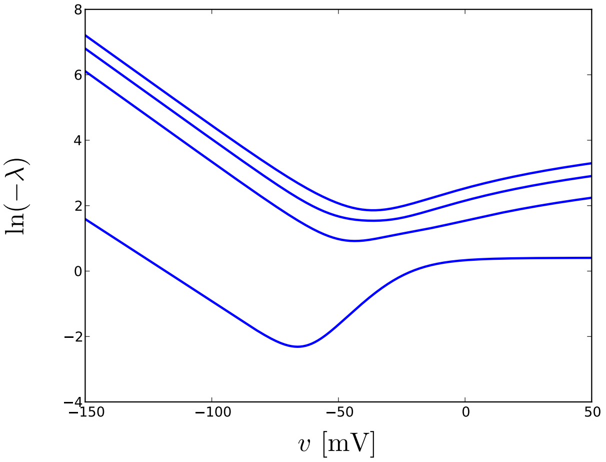


FIG. 4. Modes of the state-to-state dynamic: $\ln[-\lambda(v)]$ for $\lambda(v) \in \Lambda_{G^{(\boldsymbol{\mathcal{S}}\to\boldsymbol{\mathcal{S}}|v)}}$. All G eigenvalues are real and nonpositive. There is a zero eigenvalue associated with stationarity and four negative eigenvalues associated with decay rates from the states. Smaller $\ln(-\lambda)$ corresponds to longer time-scales. The zero eigenvalue maps to $-\infty$.

Figure 4 shows G's eigenvalues as a function of v, which indicates the voltage-dependent timescales of probability decay for modes of occupation probability. The associated decay rates play a prominent role in Fig. 5, which shows the time-dependent distribution induced over states by the 5 ms voltage pulse driving protocol.

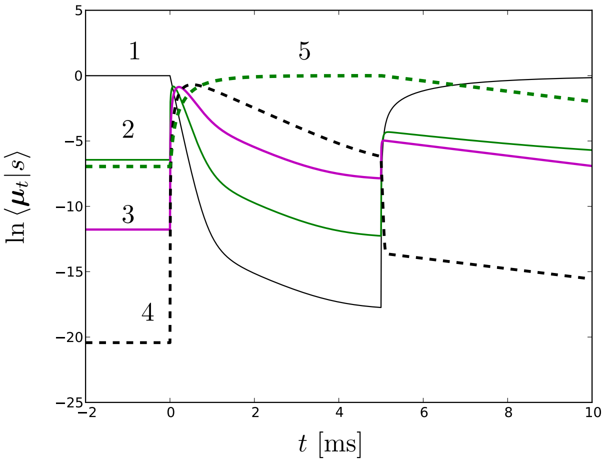


FIG. 5. Na$^+$ ion channel NESS transitions: temporal evolution of the distribution $\boldsymbol{\mu}_t$ of ion-channel conformational states induced by a deterministic 5 ms voltage pulse is shown via plots of $\ln \langle \boldsymbol{\mu}_t | s \rangle$ for all $s \in \boldsymbol{\mathcal{S}}$. Curves are labeled by the state-number to which each component corresponds.

Having the distribution over states at all times, as shown in Fig. 5, is powerful knowledge. For example, since the current through a single channel is binary—either 0 or $I_0(v)$—and since current only flows in the open conformation, the expected current through the channel is $I_0(v) = g_0[v - V_\mathrm{Na}]$ times the expectation value of being in the open state:

$$\langle I(t) \rangle = g_0 \left[v(t) - V_\mathrm{Na} \right] \langle \boldsymbol{\mu}(t) | \delta_\mathrm{open} \rangle \ ,$$

where $\delta_\mathrm{open} = (0, 0, 0, 1, 0)$, g_0 is the conductance of an open Na$^+$ channel, and:

$$V_\mathrm{Na} = \frac{k_\mathrm{B} T}{e^+} \ln \frac{[\mathrm{Na}^+]_\mathrm{out}}{[\mathrm{Na}^+]_\mathrm{in}} \approx 90 \text{ mV}$$

is the Nernst potential for sodium in a typical mammalian neuron [63].

To be clear $\langle I(t) \rangle$ is what would be observed from an ensemble of channels in a local patch of cell membrane experiencing the same driving. The current produced from the Markovian model appears to be more realistic than what would be expected from the Hodgkin–Huxley model [63]. Moreover, using our spectral-decomposition methods for functions of a Markov chain [69, 70], this current can now be obtained in closed-form.

Let us start the thermodynamic investigation by considering excess work W_ex. With $\tau = N\Delta t$, we take the limit of $\Delta t \to 0$ while keeping the product $N\Delta t = \tau$ constant. Then the expected excess work per $k_\mathrm{B}T$, from time t_0 to time $t_0 + \tau$, is:

$$\beta \langle W_{\text{ex}} \rangle = \int_{t_0}^{t_0+\tau} \langle \boldsymbol{\mu}_t | d\phi_{v(t)}/dt \rangle \; dt \; .$$

However, it should be clear that, for this stepped voltage protocol, excess work is *only performed on this system at the very onset and subsequently at the end* of the step driving. Indeed, this is the only time that the driving $v(t)$ changes and, thus, the only time that the state-dependent rate of work $d \left| \phi_{v(t)} \right\rangle / dt$ is nonzero. As we let $\Delta t \to 0$, the expected excess work (divided by $k_{\text{B}}T$) near the onset of driving becomes a step function with height:

$$\begin{aligned} &\lim_{\epsilon \to 0^+} \langle \Omega(t = \epsilon) - \Omega(t = -\epsilon) \rangle \\ &\quad = \sum_{s \in \boldsymbol{\mathcal{S}}} \langle \boldsymbol{\pi}_{\boldsymbol{v}_{\text{a}}} | s \rangle \left[\phi(10\text{mV}, s) - \phi(-100\text{mV}, s) \right] \; , \end{aligned}$$

where $\langle \boldsymbol{\pi}_{\boldsymbol{v}_{\text{a}}} | s \rangle = \pi_{-100 \text{ mV}}(s)$.

Indeed, for this singular event, the full distribution of work performed can be given according to the probabilities that the system was in a particular state when the driving was applied. For $0 < t < 5$ ms, the probability density function for βW_{ex} is:

$$\begin{aligned} p(\Omega) = &\sum_{s \in \boldsymbol{\mathcal{S}}} \pi_{-100 \text{ mV}}(s) \\ &\times \delta\Big(\Omega - \left[\phi(10\text{mV}, s) - \phi(-100\text{mV}, s) \right] \Big) \; , \end{aligned}$$

where $\delta(\cdot)$ here is the Dirac delta function. For $t > 5$ ms, the full excess environmental entropy production (EEEP) probability density function (pdf) is:

$$\begin{aligned} p(\Omega) = &\sum_{s, s' \in \boldsymbol{\mathcal{S}}} \langle \boldsymbol{\pi}_{\boldsymbol{v}_{\text{a}}} | s \rangle \langle s | \, e^{5 G_{\text{b}}} \, | s' \rangle \\ &\times \delta\Big(\Omega - \left[\phi(v_{\text{b}}, s) - \phi(v_{\text{a}}, s) \right] - \left[\phi(v_{\text{a}}, s') - \phi(v_{\text{b}}, s') \right] \Big). \end{aligned}$$

From the Dirac delta function's argument and the sum over s and s', it is clear that every nonzero-probability EEEP value Ω also has a nonzero probability for the negative $-\Omega$ of that EEEP value.

For the time-symmetric 5 ms voltage-pulse driving, Eq. (29) tells us that:

$$\frac{\Pr(\Omega, \Psi)}{\Pr(-\Omega, -\Psi)} = e^{\Psi} e^{\Omega} \; .$$

Since there are infinitely many Ψ values to account for, we do not plot the joint distribution explicitly. However, we can appreciate the necessity of the relationship by comparing it to the naive CFT interpretation that, for this case, suggests:

$$\Pr(\Omega) / \Pr(-\Omega) = e^{\Omega} \; .$$

Figure 6 compares these by plotting $e^{-\Omega} \Pr(\Omega)/\Pr(-\Omega)$.

Allowed values of the excess work that do *not* lie on $e^{-\Omega} \Pr(\Omega)/\Pr(-\Omega) = 1$ demonstrate deviations from the naive CFT interpretation. Since the constant-voltage steady states are nonequilibrium and, thus, not microscopically reversible—i.e., $\Psi \neq 0$ for some state paths—the naive CFT interpretation cannot be true despite the time-symmetric driving. The first lesson here is that the work dissipated in controlling a system maintained out of equilibrium depends on the built-in irreversibilities in the controllable dynamic. In this case, more work is dissipated than one would naively expect if the dynamics had detailed balance.

Perhaps the most surprising feature in Fig. 6 is that many of the probability ratios still *do* (almost) fall on the naive CFT line at unity. In part, this is due to a subset of the cycles in the NESS dynamic obeying detailed balance. Another contributing factor is that longer durations τ of fixed v induces a *net* dynamic $e^{\tau G}$ that *approaches* a detailed-balanced dynamic. That the values in Fig. 6 are

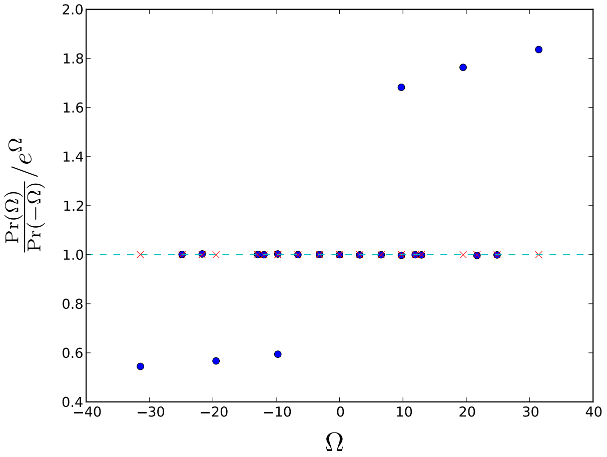


FIG. 6. Deviations from and agreement with the Crooks Fluctuation Theorem for nondetailed balance dynamics: Exact calculation of $\frac{\Pr(\Omega)}{\Pr(-\Omega)}/e^{\Omega}$ (blue dots) for all allowed values of $\Omega = \beta W_{\text{ex}}$ during the pulse drive. Since the system starts in equilibrium and since the driving is time-symmetric, a naive CFT interpretation suggests that all values lie at unity, dashed line (blue) and marked with a × (red) wherever an allowed excess work value appears. Interestingly, many of the allowed work values *do* still fall on or very near unity. Absent detailed balance, though, Eq. (28) must be used to account for the actual distribution of excess environmental entropy production and path irreversibilities that, in addition to all other values, yields the six deviant markings (blue dots) above and below unity.

sensible can be verified by checking the ratio of the joint probabilities $\langle \boldsymbol{\pi}_{\boldsymbol{v}_\mathrm{a}} | s \rangle \langle s | e^{5G_\mathrm{b}} | s' \rangle$ to the value of the joint probability with s and s' swapped.

In stark contrast to the instantaneous work contribution just analyzed, the system's excess heat Q_ex unfolds over time, exhibiting a rich structure governed by the trajectories through the conformational state-space. Recall that $-Q_\mathrm{ex}$ is the amount of work that has *actually* dissipated as heat, whereas W_ex is the work that will *potentially* be lost to heat and system entropy (if the system is allowed to relax to steady state). The expected excess heat per $k_\mathrm{B}T$ is:

$$\beta \langle Q_\mathrm{ex} \rangle = \int_{t_0}^{t_0+\tau} \langle \dot{\boldsymbol{\mu}} | \phi_{v(t)} \rangle \; dt \; . \tag{58}$$

over a duration τ, if starting at time t_0. Since $v(t)$ is constant except at the two instants of change, the integral is easily solved exactly using the fundamental theorem of calculus and Eq. (57).

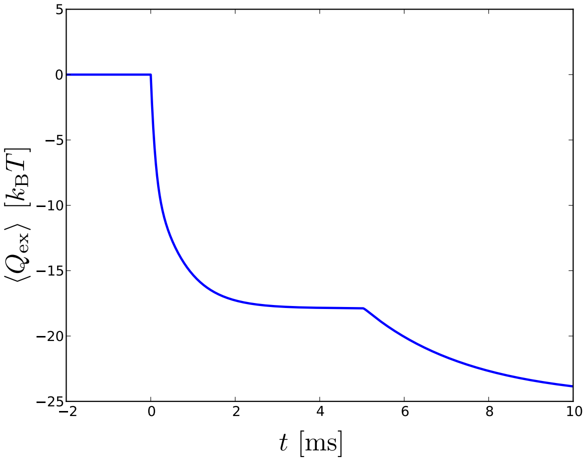


FIG. 7. Excess heat Q_ex per ion channel for an ensemble of ion channels embedded in a local patch of cell membrane; in units of $k_\mathrm{B}T \approx 26$ meV. The two bouts of relaxation correspond to the ion channel adapting to sudden changes in voltage across the cell membrane.

For $t_0 < t_0 + \tau < 0$, the system is in the initial steady-state and has a time-independent heat and so a constant excess heat rate that vanishes:

$$\frac{\langle Q_\mathrm{ex} \rangle}{\tau} = 0 \; .$$

Figure 7 shows the expected excess heat $\langle Q_\mathrm{ex} \rangle$ over the course of the voltage-drive protocol. The steady-state average rate of excess heat production within any steady state is necessarily zero. However, the channel macromolecule responds to changes in the environment via conformational changes and corresponding heat productions that unfold on the timescale of milliseconds [71]. Notably, the expected heat is on the order of tens of $k_\mathrm{B}T$, which for mammalian neurons is $k_\mathrm{B}T = 1/\beta \approx 26$ meV.

For $0 = t_0 < t_0 + \tau < 5$ ms, the system has a time-varying heat production as it synchronizes to the new nonequilibrium steady state. We see that the applied work is dissipated as heat over the time scale of milliseconds. The simplicity of both the system and driving protocol provide a good opportunity to review the meaning and behavior of the excess thermodynamic quantities during this epoch. From Eq. (57), we know that $\langle \boldsymbol{\mu}_t | = \langle \boldsymbol{\pi}_{v_\mathrm{a}} | e^{tG_\mathrm{b}}$ in this case. At the same time, the set of possible steady-state-surprisals $|\phi_{v(t)}\rangle$ is time-independent over the set of conformational states during this epoch since v is temporarily fixed at $v_\mathrm{b} = 10$ mV. Since the state-occupation is time-dependent, however, the steady-state surprisal $\phi(v_\mathrm{b}, s_t)$ and its expected value $\langle \phi \rangle = \langle \boldsymbol{\mu}_t | \phi_{v(t)} \rangle$ are nevertheless time-dependent in a way that offsets Q_ex, yielding a constant $\Omega = \Delta\phi - \beta Q_\mathrm{ex}$ during this epoch of relaxation. We note that the net dissipated work W_diss varies over this interval according not only to Q_ex but also according to the time-varying state entropy h. W_diss asymptotically approaches W_ex during any period of relaxation.

Again, as seen in Fig. 7, the expected excess heat drops for several milliseconds after the final voltage switch at 5 ms, as the ion channel re-adapts to its original steady state. For this second bout of relaxation the adaptation is slower since, in accordance with Fig. 4, the slowest timescale at $v_\mathrm{a} = -100$ mV is slower than the slowest timescale at $v_\mathrm{b} = 10$ mV.

Overall, we see that the excess thermodynamic quantities are well behaved and accessible. In particular, without needing to know the background biological upkeep of the Na$^+$ ion channel, we can access and control coarse degrees of freedom of the channel macromolecule via modulating the voltage across the cell membrane. Moreover, for the Na$^+$ channel, state-measurement and feedback could be implemented on the timescale of milliseconds before the system has fully relaxed back to steady state while dissipating on the order of tens of $k_\mathrm{B}T$ per adaptive response per ion channel. Fortuitously, this suggests an accessible platform for laboratory experimentation where dissipation during partial control can be further explored. Next, we comment briefly on an intrinsic type of measurement and feedback that happens in vivo every moment.

D. Intrinsic feedback

Having come this far, we close illustrating the thermodynamics of NESS transitions with a final application. In a biologically active (in vivo) neuron, the input membrane voltage at each time depends on integrated current—a functional of the state distribution—up to that time. Section V F's relations for modified integral fluctuation relations describe the thermodynamic agency of Na$^+$ channels in vivo, whereas conventional fluctuation relations fall short. Although there is certainly feedback in vivo, it is not the "feedback control" discussed recently. Importantly, no "outsider" forces the feedback; the feedback is intrinsic—woven into the system–environment joint dynamic. We leave a thorough investigation of the thermodynamics of intrinsic feedback to elsewhere. The success here, however, already suggests investigating other natural systems with intrinsic feedback—in joint nonlinear dynamics and complex networks—to test the new fluctuation theorems and computational methods in a broader class of interacting complex nonequilbrium systems.

VIII. DISCUSSION

In light of our refined detailed fluctuation theorem Eq. (28) for nondetailed-balanced dynamics, we referred to the belief in Eq. (32) as the naive CFT interpretation since it appeals to a nonphysical conjugate dynamics, as in Eq. (30). Similarly, we referred to failures of Eq. (31) as CFT violations. Nonetheless, with proper interpretation using the unphysical conjugate dynamics, Eq. (30) is mathematically correct even without detailed balance and can be a useful device for establishing integral FTs.

We hope that our nonintegral FTs—especially Eq. (28) that constrains the joint distribution of excess and housekeeping entropies—provide better physical intuition for the structure of effective dynamics outside detailed balance. Similarly, we hope that Eq. (54) leads to new opportunities in which driven systems are controlled more efficiently by harnessing the system's intrinsic irreversibilities. This could be key, to mention one example, in influencing self-assembly of active matter.

Path irreversibility clearly plays a prominent role in each of these results. Although the preceding introduced a unifying framework, certain subclasses of path irreversibility have already been proposed recently, as we noted. In certain applications, path irreversibility is governed by differences in chemical potential. In such cases, irreversibility is quantitatively related to cycle affinities. For example, Ref. [72] discovered a special case of the results developed here specifically applicable to the interesting example of a kinesin motor protein.

Along similar lines, Ref. [73] recently elaborated one type of irreversibility, called *absolute irreversibility*, that at first appears to constitute an extreme contribution to the total path irreversibility Ψ. This indeed is one interpretation, but not the full story. On closer examination, this result appears to coincide most directly with Eq. (38), which must be used when starting in a nonsteady-state. It does not coincide with a violation of Eq. (39), which is simply inapplicable when starting in a nonsteady-state. And so, we reinterpret this result as emphasizing the importance of the nonsteady-state contribution to free energy change, $\beta^{-1}\Delta(\gamma)$. Explicitly:

$$\begin{aligned} \left\langle e^{-\beta W_{\text{diss}}} \right\rangle_{\Pr(s_{0:N}|\boldsymbol{\mu}_{\mathbf{F}}, \mathbf{x})} &= 1 \\ &\Rightarrow \langle \Omega \rangle \geq \Delta \langle \gamma \rangle \ . \end{aligned}$$

This captures, for example, the entropy change associated with free expansion. From our viewpoint, however, any absolute irreversibility is only one extreme of the broader generalization introduced above to explore the consequences of irreversibility and nonsteady-state additions to free energy.

To frame our results in yet another way, we note that the "feedback control" imposed by an experimenter on an otherwise detailed-balanced system can produce an apparent nondetailed-balanced dynamic, if one fails to account for the feedback controller. Oddly, the thermodynamic implications of this artificial sort of feedback has garnered more attention recently than the intrinsic feedback that fuels complexity all around us. Living systems are the true flagship of complex physical agents with intrinsic computational feedback across many levels of their organization. In principle, our fluctuation relations describe all of these aspects, together. Yet, there is much still to learn as we unravel the thermodynamic consequences of intrinsic feedback.

In particular, exploring the theory developed here may suggest how a system's internal model of its environment, together with an action policy that leverages knowledge captured in the model to control the environment, allows the system to play the survival game to its thermodynamic advantage. For example, an agent can use information about the environment to increase its nonsteady-state free energy and perform useful work—a phenomenon that is not only reminiscent of living beings, but also comes very near to defining them.

A sequel to the above derives exact closed-form expressions for the moments of excess heat and excess work when the joint system–environment dynamic is governed by a (finite or countably infinite) discrete- or continuous-time hidden Markov model. A joint system can always be modeled as a joint hidden Markov model—at least as an approximation to the true joint dynamics. For this reason, our exact results should provide broadly applicable tools. The latter have particular theoretical advantage in giving access to what occurs in transient and asymptotic dynamics of excess thermodynamic quantities atop

NESSs.

We hope that our results and methods stimulate investigating the excess thermodynamics of systems with intrinsic feedback—from designed "toy demons" to complex biological molecules affected by and simultaneously affecting their environments. Several biological examples that suggest themselves include kinesin motors [74], drug-operated channels [75], and dynamic synapses [76], just to name a few.

IX. CONCLUSION

We presented generalized fluctuation theorems for complex nonequilibrium systems driven between NESSs. In addition to the detailed FTs that constrain joint distributions of excess and housekeeping quantities, we introduced integral fluctuation theorems in the presence of an auxiliary variable. The auxiliary variable need not be measurement nor any other meddling of an outsider. Due to this, it generalizes the theory of "feedback control" to the setting of arbitrary intrinsic feedback between system and environment.

In summary, the traditional laws of thermodynamics are largely preserved for the renormalized "excess" thermodynamic quantities that arise naturally when considering nondetailed-balanced complex systems. However, the laws must be modified by the entropic contribution of path irreversibility. We noted that the latter turns out to be equivalent to steady-state thermodynamics' housekeeping entropy production.

We discussed the operational meanings of the excess and dissipated quantities, in contradistinction to the housekeeping entropy production. They are more than mere numbers that can be calculated. Indeed, they are distinguished in the "real world" as energies that must be exerted for practical control of complex systems. More specifically, we showed that the excess and dissipated works are important since they describe the energetic cost of controlling an active system that is kept out of equilibrium by any number of uncontrolled processes; e.g., by photosynthesis or oxidative phosphorylation. Moreover, we identified an opportunity to leverage the intrinsic irreversibilities in an active system to influence the state while exerting less energy than would be required to influence a nonactive system. We also demonstrated the importance of these quantities beyond the setting of unidirectional control. For example, our generalization of feedback-control is relevant for active systems that influence each other through autonomous feedback.

Our relations still hold for excursions between equilibrium steady states, but we then have the simplification that $\Psi = \beta Q_{\text{hk}} = 0$. Consistently, equilibrium thermodynamics is a reduction of the theory of excess thermodynamic quantities with no housekeeping terms—when all paths are microscopically reversible.

Layers of emergence, typical of the biological world [77, Fig. 6], beg renormalization in terms of a hierarchy of housekeeping backgrounds [78]. The opportunity offered up by emergent levels of novel organization is a new richness in nondetailed-balanced effective dynamics—dynamics and structure that can be exploited by intelligent thermodynamic agency [79, 80]. We consider the thermodynamics of agency in a sequel, analyzing a simple autonomous agent that harvests energy by leveraging hidden correlations in a fluctuating environment.

We leave the development for now, but with an encouraging lesson: Even in nonstationary nonequilibrium, there is excess thermodynamic structure at any level of observation that we can access, control, and harness.

ACKNOWLEDGMENTS

We thank Tony Bell, Alec Boyd, Gavin Crooks, Sebastian Deffner, Chris Jarzynski, John Mahoney, Dibyendu Mandal, and Adam Rupe for useful feedback. We thank the Santa Fe Institute for its hospitality during visits. JPC is an SFI External Faculty member. This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-12-1-0234, W911NF-13-1-0390, and W911NF-13-1-0340.

Appendix A: Extension to Non-Markovian Instantaneous Dynamics

Commonly, theoretical developments assume state-to-state transitions are instantaneously Markovian given the input. This assumption works well for many cases, but fails in others with strong coupling between system and environment. Fortunately, we can straightforwardly generalize the results of stochastic thermodynamics by considering a system's observable states to be functions of latent variables $\boldsymbol{\mathcal{R}}$. The goal in the following is to highlight the necessary changes, so that it should be relatively direct to adapt our derivations to the non-Markovian dynamics. See Ref. [81] for an alternative approach to addressing non-Markovian dynamics.

a. Latent states, system states, and their many distributions

Even with constant environmental input, the dynamic over a system's states need not obey detailed balance nor exhibit any finite Markov order. We assume that the classical observed states $\boldsymbol{\mathcal{S}}$ are functions $f : \boldsymbol{\mathcal{R}} \to \boldsymbol{\mathcal{S}}$ of a latent Markov chain. We also assume that the stochastic transitions among latent states are determined by the current environmental input $x \in \mathcal{X}$, which can depend arbitrarily on all previous input and system-state history. The Perron–Frobenius theorem guarantees that there is a stationary distribution over latent states associated with each fixed input x; the function of the Markov chain maps this stationary distribution over latent states into the stationary distribution over system states. These are the stationary distributions associated with system NESSs.

We assume too that the $\boldsymbol{\mathcal{R}}$-to-$\boldsymbol{\mathcal{R}}$ transitions are Markovian given the input. However, different inputs induce different Markov chains over the latent states. This can be described by a (possibly infinite) set of input-conditioned transition matrices over the latent state set $\boldsymbol{\mathcal{R}}$: $\{\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x)}\}_{x\in\mathcal{X}}$, where $\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x)}_{i,j} = \Pr(\mathcal{R}_t = r^j|\mathcal{R}_{t-1} = r^i, X_t = x)$. Probabilities regarding actual state paths can be obtained from the latent-state-to-state transition dynamic together with the observable-state projectors, which we now define.

We denote distributions over the latent states as bold Greek symbols, such as $\boldsymbol{\mu}$. As in the main text, it is convenient to cast $\boldsymbol{\mu}$ as a row-vector, in which case it appears as the bra $\langle\boldsymbol{\mu}|$. The distribution over latent states $\boldsymbol{\mathcal{R}}$ implies a distinct distribution over observable states $\boldsymbol{\mathcal{S}}$. A sequence of driving inputs updates the distribution: $\boldsymbol{\mu}_{t+n}(\boldsymbol{\mu}_t, x_{t:t+n})$. In particular:

$$\begin{aligned}\langle\boldsymbol{\mu}_{t+n}| &= \langle\boldsymbol{\mu}_t|\,\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x_{t:t+n})} \\ &= \langle\boldsymbol{\mu}_t|\,\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x_t)}\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x_{t+1})}\ldots\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x_{t+n-1})} ~.\end{aligned}$$

(Recall that time indexing is denoted by subscript ranges $n : m$ that are left-inclusive and right-exclusive.) An infinite driving history $\vec{x}$ induces a distribution $\boldsymbol{\mu}(\vec{x})$ over the state space, and $\boldsymbol{\pi_x}$ is the specific steady-state distribution over $\boldsymbol{\mathcal{R}}$ induced by tireless repetition of the single environmental drive x. Explicitly:

$$\langle\boldsymbol{\pi_x}| = \lim_{n\to\infty}\langle\boldsymbol{\mu}_0|\left(\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x)}\right)^n ~.$$

Usefully, $\boldsymbol{\pi_x}$ can also be found as the left eigenvector of $\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x)}$ associated with the eigenvalue of unity:

$$\langle\boldsymbol{\pi_x}| = \langle\boldsymbol{\pi_x}|\,\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x)} ~. \tag{A1}$$

The physically relevant steady-state probabilities are this vector's projection onto observable states: $\pi_x(s) = \langle\boldsymbol{\pi_x}|s\rangle$, where $|s\rangle = |\delta_{s,f(r)}\rangle$ has a vector-representation in the latent-state basis with elements of all 0s except 1s where the latent state maps to the observable state s.

Assuming latent-state-to-state transitions are Markovian allows the distribution $\boldsymbol{\mu}$ over these latent states to summarize the causal relevance of the entire driving history.

b. Implications

A semi-infinite history induces a particular distribution over system latent states and implies another particular distribution over its observable states. This can be usefully recast in terms of the "start" (or initial) distribution $\boldsymbol{\mu}_0$ induced by the path $x_{-\infty:1}$ and the driving history $x_{1:t+1}$ since then, giving the entropy of the induced state distribution:

$$\begin{aligned}h^{(s|\boldsymbol{\mu}_0,x_{1:t+1})} &= -\ln\Pr(\mathcal{S}_t = s|\boldsymbol{\mu}_0, x_{1:t+1}) \\ &= -\ln\langle\boldsymbol{\mu}_0|\,\mathsf{T}^{(\boldsymbol{\mathcal{R}}\to\boldsymbol{\mathcal{R}}|x_{1:t+1})}\,|s\rangle ~.\end{aligned}$$

Or, employing the new distribution and the driving history since then, the path entropy (functional of state and driving history) can be expressed simply in terms of the *current* distribution over latent states and the candidate observable state s:

$$\begin{aligned}h^{(s|\boldsymbol{\mu})} &= -\ln\Pr(\mathcal{S}_t = s|\mathcal{R}_t \sim \boldsymbol{\mu}) \\ &= -\ln\langle\boldsymbol{\mu}|s\rangle ~.\end{aligned}$$

Averaging the path-conditional state entropy over observable states again gives a genuine input-conditioned Shannon state entropy:

$$\langle h^{(s_t|\overleftarrow{x}_t)}\rangle_{\Pr(s_t|\overleftarrow{x}_t)} = \mathrm{H}[\mathcal{S}_t|\overleftarrow{X}_t = \overleftarrow{x}_t] ~.$$

It is again easy to show that the state-averaged path entropy $k_\mathrm{B}\,\mathrm{H}[\mathcal{S}_t|\overleftarrow{x}_t]$ is an extension of the system's steady-state nonequilibrium entropy. In steady-state, the state-averaged path entropy reduces to:

$$\begin{aligned}k_\mathrm{B}\,\mathrm{H}[\mathcal{S}_t|\overleftarrow{X}_t = \ldots xxx] &= -k_\mathrm{B}\,\mathrm{H}[\mathcal{S}_t|\mathcal{R}_t \sim \boldsymbol{\pi_x}] \\ &= -k_\mathrm{B}\sum_{s\in\boldsymbol{\mathcal{S}}}\pi_x(s)\ln\pi_x(s) \\ &= S_\mathrm{ss}(x) ~.\end{aligned}$$

The *nonsteady-state addition to free energy* is:

$$\beta^{-1}\gamma(s|\boldsymbol{\mu}, x) \equiv \beta^{-1}\ln\frac{\Pr(\mathcal{S}_t = s|\mathcal{R}_{t-1} \sim \boldsymbol{\mu}, X_t = x)}{\pi_x(s)} ~.$$

Averaging over observable states this becomes the relative entropy:

$$\langle \gamma(s|\boldsymbol{\mu}, x) \rangle = D_{\mathrm{KL}} \left[\Pr(\mathcal{S}_t | \mathcal{R}_{t-1} \sim \boldsymbol{\mu}, X_t = x) || \boldsymbol{\pi}_{\boldsymbol{x}} \right] ~,$$

which is always nonnegative.

Using this setup and decomposing:

$$\frac{\Pr(\mathcal{S}_{0:N} = s^0 \mathbf{s} | \mathcal{R}_{-1} \sim \boldsymbol{\mu}_{\mathrm{F}}, X_{0:N} = x^0 \mathbf{x})}{\Pr(\mathcal{S}_{0:N} = s^{N-1} \mathbf{s}^{\mathrm{R}}_{\leftarrow} | \mathcal{R}_{-1} \sim \boldsymbol{\mu}_{\mathrm{R}}, X_{0:N} = x^N \mathbf{x}^{\mathrm{R}})}$$

in analogy with Eq. (25), it is straightforward to extend the remaining results of the main body to the setting in which observed states are functions of a Markov chain. Notably, the path dependencies pick up new contributions from non-Markovity. Also, knowledge of distributions over latent states provides a thermodynamic advantage to Maxwellian Demons.

Appendix B: Integral fluctuation theorems with auxiliary variables

Recall that we quantify how much the auxiliary variable independently informs the state sequence via the nonaveraged conditional mutual information:

$$\begin{aligned} i[\overrightarrow{s}; \overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}] &\equiv \ln \frac{\Pr(\overrightarrow{s}, \overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}) \Pr(\overrightarrow{s} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}})} \\ &= \ln \frac{\Pr(\overrightarrow{s}, \overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y}, \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}) \Pr(\overrightarrow{s} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}})} ~. \end{aligned}$$

Note that averaging over the input, state, and auxiliary sequences gives the familiar conditional mutual information:

$$\begin{aligned} &\mathrm{I}[\mathcal{S}_{0:N}; Y_{0:N} | X_{0:N}, \boldsymbol{\mu}_{\mathrm{F}}] \\ &\qquad = \langle i[\overrightarrow{s}; \overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}] \rangle_{\Pr(x_{0:N}, s_{0:N}, y_{0:N} | \boldsymbol{\mu}_{\mathrm{F}})} ~. \end{aligned}$$

(Averaging over distributions is the same as being given the distribution, since the distribution over distributions is assumed to be peaked at $\boldsymbol{\mu}_{\mathrm{F}}$.)

Noting that:

$$\begin{aligned} &e^{\beta W_{\mathrm{diss}} + i(\overrightarrow{s}; \overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}) + \Psi} \\ &\qquad = e^{\Omega + i(\overrightarrow{s}; \overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}) + \Psi + (\gamma_{\mathrm{F}} - \gamma_{\mathrm{R}})} \\ &\qquad = \frac{\Pr(\overrightarrow{s}, \overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \Pr(s^{N-1} \mathbf{s}^{\mathrm{R}}_{\leftarrow} | \mathbf{x}^{\mathrm{R}} x^0, \boldsymbol{\mu}_{\mathrm{R}})} \\ &\qquad = \frac{\Pr(\overrightarrow{s}, \overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}})}{\Pr(\overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \Pr(\overleftarrow{s} | \overleftarrow{x}, \boldsymbol{\mu}_{\mathrm{R}})} ~, \end{aligned}$$

where $\boldsymbol{\mu}_{\mathrm{R}} = \boldsymbol{\mu}(\boldsymbol{\mu}_{\mathrm{F}}, \vec{x})$, we have the integral fluctuation theorem (IFT):

$$\begin{aligned} &\left\langle e^{-\beta W_{\mathrm{diss}} - i(\overrightarrow{s}; \overrightarrow{y} | \overrightarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}) - \Psi} \right\rangle_{\Pr(\overrightarrow{s}, \overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}})} \\ &\quad = \sum_{\overrightarrow{x}, \overrightarrow{s}, \overrightarrow{y}} \Pr(\overrightarrow{s}, \overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \frac{\Pr(\overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \Pr(\overleftarrow{s} | \overleftarrow{x}, \boldsymbol{\mu}_{\mathrm{R}})}{\Pr(\overrightarrow{s}, \overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}})} \\ &\quad = \sum_{\overrightarrow{x}, \overrightarrow{s}, \overrightarrow{y}} \Pr(\overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \Pr(\overleftarrow{s} | \overleftarrow{x}, \boldsymbol{\mu}_{\mathrm{F}}) \\ &\quad = \sum_{\overrightarrow{x}, \overrightarrow{y}} \Pr(\overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \sum_{\overleftarrow{s}} \Pr(\overleftarrow{s} | \overleftarrow{x}, \boldsymbol{\mu}_{\mathrm{R}}) \\ &\quad = \sum_{\overrightarrow{x}, \overrightarrow{y}} \Pr(\overrightarrow{y}, \overrightarrow{x} | \boldsymbol{\mu}_{\mathrm{F}}) \\ &\quad = 1 ~. \end{aligned}$$

Notably, this relation holds arbitrarily far from equilibrium and allows for the starting and ending distributions to both be nonsteady-state.

It is tempting to conclude that the revised Second Law of Thermodynamics should read:

$$\langle W_{\mathrm{diss}} \rangle \geq -k_{\mathrm{B}} T ~ \mathrm{I}[\vec{\mathcal{S}}; \vec{Y} | \vec{X}, \boldsymbol{\mu}_{\mathrm{F}}] - \langle Q_{\mathrm{hk}} \rangle ~, \qquad \text{(B1)}$$

which includes the effects of both irreversibility and conditional mutual information between state-sequence and auxiliary sequence, given input-sequence. However, we expect that $\langle Q_{\mathrm{hk}} \rangle > 0$, so Eq. (B1) is not the strongest bound derivable. Dropping Ψ from the IFT still yields a valid equality. However, the derivation runs differently since it depends on the normalization of the dual dynamic: quantities of the form:

$$\frac{\pi_{x^n}(s^{n-1})}{\pi_{x^n}(s^n)} \Pr(s^n | s^{n-1}, x^n) ~.$$

These are mathematically-sound transition probabilities $\widetilde{\mathsf{T}}^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}} | x^n)}_{s^n, s^{n-1}}$, but only of a nonphysical artificial dynamic. Although IFTs with Ψ may be useful for other reasons, it is the non-Ψ IFTs that seem to yield the tighter bound for the revised Second Laws of information thermodynamics without detailed balance.

[1] G. E. Crooks. On thermodynamic and microscopic reversibility. *J. Stat. Mech.: Th. Exp.*, 2011(07):P07008, 2011. 1

[2] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.*, 90(5/6):1481–1487, 1998. 1, 9, 12

[3] T. Sagawa and M. Ueda. Nonequilibrium thermodynamics of feedback control. *Phys. Rev. E*, 85:021104, Feb 2012. 1, 3, 15

[4] H. Wang and G. Oster. Energy transduction in the F1 motor of ATP synthase. *Nature*, 396(6708):279–282, 1998. 1

[5] M. Polettini and M. Esposito. Irreversible thermodynamics of open chemical networks. I. Emergent cycles and broken conservation laws. *J. Chem. Physics*, 141(2), 2014. 1

[6] R. Landauer. Statistical physics of machinery: Forgotten middle-ground. *Physica A: Stat. Mech. App.*, 194(1-4):551–562, 1993. 1

[7] H. Qian. Nonequilibrium steady-state circulation and heat dissipation functional. *Phys. Rev. E*, 64:022101, 2001. 1

[8] W. Horsthemke. Noise induced transitions. In C. Vidal and A. Pacault, editors, *Non-Equilibrium Dynamics in Chemical Systems: Proceedings of the International Symposium, Bordeaux, France, September 3–7, 1984*, pages 150–160, Berlin, Heidelberg, 1984. Springer. 1

[9] B. Lindner, J. Garcia-Ojalvo, A. Neiman, and L. Schimansky-Geier. Effects of noise in excitable systems. *Physics Reports*, 392(6):321 – 424, 2004.

[10] J. P. Crutchfield and C. Aghamohammdi. Not all fluctuations are created equal: Spontaneous variations in thermodynamic function. arxiv.org:1609.02519. 1

[11] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994. 1, 3

[12] P. M. Riechers. *Exact Results Regarding the Physics of Complex Systems via Linear Algebra, Hidden Markov Models, and Information Theory.* PhD thesis, University of California, Davis, 2016. 1, 16

[13] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Physics*, 75(12):126001, 2012. 2

[14] R. Spinney and I. Ford. Fluctuation relations: A pedagogical overview. In *Nonequilibrium Statistical Physics of Small Systems*, pages 3–56. Wiley-VCH Verlag GmbH & Co. KGaA, 2013. 2

[15] Y. Oono and M. Paniconi. Steady state thermodynamics. *Prog. Theo. Phys. Supp.*, 130:29–44, 1998. 2

[16] T. Hatano and S. Sasa. Steady-state thermodynamics of Langevin systems. *Phys. Rev. Lett.*, 86:3463–3466, 2001. 3, 5, 7, 13, 14

[17] E. H. Trepagnier, C. Jarzynski, F. Ritort, G. E. Crooks, C. J. Bustamante, and J. Liphardt. Experimental test of Hatano and Sasa's nonequilibrium steady-state equality. *Proc. Natl. Acad. Sci. USA*, 101(42):15038–15041, 2004. 2

[18] D. Mandal and C. Jarzynski. Analysis of slow transitions between nonequilibrium steady states. *J. Stat. Mech.: Th. Exp.*, 2016(6):063204, 2016. 2, 5, 6, 7

[19] D. J. Evans, D. J. Searles, and S. R. Williams. The Evans–Searles fluctuation theorem. In *Fundamentals of Classical Statistical Thermodynamics*, pages 49–64. Wiley-VCH Verlag GmbH & Co. KGaA, 2016. 2

[20] D. J. Evans, E.G.D. Cohen, and G.P. Morriss. Probability of second law violations in shearing steady states. *Physical review letters*, 71(15):2401, 1993. 2

[21] G. Gallavotti and E.G.D. Cohen. Dynamical ensembles in stationary states. *J. Stat. Phys.*, 80(5):931–970, 1995. 5

[22] D. J. Evans, D. J. Searles, and L. Rondoni. Application of the gallavotti-cohen fluctuation relation to thermostated steady states near equilibrium. *Phys. Rev. E*, 71(5):056120, 2005. 2

[23] M. Esposito and C. Van den Broeck. Three detailed fluctuation theorems. *Phys. Rev. Lett.*, 104:090601, 2010. 2, 5, 6, 13

[24] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60(3):2721–2726, 1999. 2, 6, 9, 12

[25] E. Roldán and J. M. R. Parrondo. Estimating dissipation from single stationary trajectories. *Physical Rev. Let.*, 105(15):150607, 2010. 3

[26] J. M. Horowitz and S. Vaikuntanathan. Nonequilibrium detailed fluctuation theorem for repeated discrete feedback. *Phys. Rev. E*, 82:061120, Dec 2010. 3

[27] J. L. England. Statistical physics of self-replication. *J. Chem. Physics*, 139(12):–, 2013. 3, 16, 18

[28] J. L. England. Dissipative adaptation in driven self-assembly. *Nature Nanotech.*, 10(11):919–923, 2015. 18

[29] N. Perunov, R. A. Marsland, and J. L. England. Statistical physics of adaptation. *Phys. Rev. X*, 6(2):021036, 2016. 3, 16, 18

[30] We ignore nonergodicity to simplify the development. The approach, though, handles nonergodicity just as well. However, distracting nuances arise that we do not wish to dwell on. For example, if the Markov chain has more than one attracting component for a particular x, then $\boldsymbol{\pi_x}$ is not unique, but can be constructed as any one of infinitely many probability-normalized linear superpositions of left eigenvectors of $\mathsf{T}^{(\boldsymbol{\mathcal{S}} \to \boldsymbol{\mathcal{S}}|x)}$ associated with the eigenvalue of unity. 4

[31] D. Ruelle and F. Takens. On the nature of turbulence. *Comm. Math. Physics*, 20(3):167–192, 1971. 4

[32] M. C. Mackey. *Time's arrow: The origins of thermodynamic behavior.* Courier Corporation, 2003. 4

[33] We start in a discrete-time setup, but later translate to continuous time. 5

[34] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, second edition, 2006. 5, 8

[35] The sign conventions adopted for Q, Q_{hk}, and Q_{ex} are slightly inharmonious. We take Q and Q_{ex} to be energy that spontaneously flows *into* a system at fixed x, whereas we have chosen for Q_{hk} to have the opposite sign convention, for easy comparison to the literature. As a result, our quantities technically satisfy $Q_{\text{ex}} = Q + Q_{\text{hk}}$,

rather than $Q_{\mathrm{ex}} = Q - Q_{\mathrm{hk}}$. 6

[36] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks. Thermodynamics of prediction. *Phys. Rev. Lett.*, 109:120604, Sep 2012. 6

[37] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989. 6

[38] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu. The energy-speed-accuracy trade-off in sensory adaptation. *Nature Physics*, 8(5):422–428, 2012. 6

[39] P. Sartori, L. Granger, C. F. Lee, and J. M. Horowitz. Thermodynamic costs of information processing in sensory adaptation. *PLoS Comp. Bio.*, 10(12):e1003974, 12 2014. 15

[40] D. Hartich, A. C. Barato, and U. Seifert. Sensory capacity: an information theoretical measure of the performance of a sensor. *Phys. Rev. E*, 93(2):022116, 2016. 6

[41] M. Esposito, U. Harbola, and S. Mukamel. Entropy fluctuation theorems in driven open systems: Application to electron counting statistics. *Phys. Rev. E*, 76:031132, Sep 2007. 6

[42] G. B. Bagci, U. Tirnakli, and J. Kurths. The second law for the transitions between the non-equilibrium steady states. *Phys. Rev. E*, 87:032161, Mar 2013. 6

[43] To be more precise, we write $\Pr(\mathcal{S}_t = s | \mathcal{S}_0 \sim \boldsymbol{\mu}_0, x_{1:t+1})$ as $\Pr_{\mathcal{S}_0 \sim \boldsymbol{\mu}_0}(\mathcal{S}_t = s | x_{1:t+1})$, since the probability is not *conditioned* on $\boldsymbol{\mu}_0$—a probability measure for subsequent state sequences. Here, we simply gloss over this nuance, later adopting the shorthand: $\Pr(\mathcal{S}_t = s | \boldsymbol{\mu}_0, x_{1:t+1})$. 7

[44] B. Gaveau and L. S. Schulman. A general framework for non-equilibrium phenomena: The master equation and its formal consequences. *Phys. Lett. A*, 229(6):347–353, 1997. 8, 13

[45] D. A. Sivak and G. E. Crooks. Near-equilibrium measurements of nonequilibrium free energy. *Phys. Rev. Let.*, 108(15), 2012.

[46] S. Deffner and E. Lutz. Information free energy for nonequilibrium states. *arXiv:1201.3888*, 2012. 8

[47] H. Qian. Cycle kinetics, steady state thermodynamics and motors: A paradigm for living matter physics. *J. Physics: Cond. Matt.*, 17(47):S3783, 2005. 8

[48] Liepelt, S. and Lipowsky, R. Steady-state balance conditions for molecular motor cycles and stochastic nonequilibrium processes. *Euro. Phys. Let.*, 77(5):50002, 2007.

[49] S. Liepelt and R. Lipowsky. Kinesin's network of chemomechanical motor cycles. *Phys. Rev. Lett.*, 98:258102, Jun 2007. 8

[50] G. E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61(3):2361–2366, 2000. 12, 14

[51] V. Y. Chernyak, M. Chertkov, and C. Jarzynski. Path-integral analysis of fluctuation theorems for general Langevin processes. *J. Stat. Mech.: Th. Exp.*, 2006(08):P08001, 2006. 12, 13

[52] R. J. Harris and G. M. Schutz. Fluctuation theorems for stochastic dynamics. *J. Stat. Mech.: Th. Exp.*, 2007(07):P07020, 2007. 13

[53] U. Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Phys. Rev. Lett.*, 95:040602, Jul 2005. 13

[54] S. Lahiri and A. M. Jayannavar. Fluctuation theorems for excess and housekeeping heat for underdamped Langevin systems. *Euro. Phys. J. B*, 87(9), 2014. 13

[55] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997. 14

[56] T. Speck and U. Seifert. Integral fluctuation theorem for the housekeeping heat. *J. Phys. A: Math. Gen.*, 38(34):L581, 2005. 14

[57] S. Vaikuntanathan and C. Jarzynski. Dissipation and lag in irreversible processes. *Europhys. Lett.*, 87(6):60005, 2009. 14

[58] Since $e^{\langle -Y \rangle} \leq \langle e^{-Y} \rangle$ and $\ln(a)$ is monotonically increasing for positive-valued $a \in \{e^{-\langle Y \rangle}, \langle e^{-Y} \rangle\}$). 17

[59] T. O'Leary, A. H. Williams, A. Franci, and E. Marder. Cell types, network homeostasis, and pathological compensation from a biologically plausible ion channel expression model. *Neuron*, 82(4):809–821, 2014. 18

[60] G. G. Turrigiano and S. B. Nelson. Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.*, 5(2):97–107, 2004. 18

[61] B. Sengupta and M. B. Stemmler. Power consumption during neuronal computation. *Proc. IEEE*, 102(5):738–750, May 2014. 18, 19

[62] C. Howarth, C. M. Peppiatt-Wildman, and D. Attwell. The energy use associated with neural computation in the cerebellum. *J. Cerebral Blood Flow & Metabolism*, 30(2):403–414, 2010. 18

[63] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* Computational Neuroscience Series. MIT Press, Boston, Massachusetts, revised edition, 2005. 18, 19, 20, 22

[64] D. Attwell and S. B. Laughlin. An energy budget for signaling in the grey matter of the brain. *J. Cerebral Blood Flow & Metabolism*, 21(10):1133–1145, 2001. 18

[65] E. M. Izhikevich. *Dynamical Systems in Neuroscience.* Computational Neuroscience Series. MIT Press, Boston, Massachusetts, 2010. 18

[66] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code.* Bradford Books, New York, 1999. 18

[67] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physio.*, 117(4):500, 1952. 19, 20

[68] J. Patlak. Molecular kinetics of voltage-dependent Na^+ channels. *Physiol. Rev.*, 71(4):1047–1080, 1991. 19

[69] J. P. Crutchfield, C. J. Ellison, and P. M. Riechers. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2016. 20, 22

[70] P. M. Riechers and J. P. Crutchfield. Beyond the spectral theorem: Decomposing arbitrary functions of nondiagonalizable operators. 2016. arxiv.org:1607.06526 [math-ph]. 20, 22

[71] The characteristic timescale is actually the net result of a combination of timescales from the inverse eigenvalues

of G. Of necessity, these are the same timescales that determine the relaxation of the state distribution. 24

[72] D. Lacoste, A. W. C. Lau, and K. Mallick. Fluctuation theorem and large deviation function for a solvable model of a molecular motor. *Phys. Rev. E*, 78:011915, 2008. 25

[73] Y. Murashita, K. Funo, and M. Ueda. Nonequilibrium equalities in absolutely irreversible processes. *Phys. Rev. E*, 90:042110, Oct 2014. 25

[74] B. Altaner, A. Wachtel, and J. Vollmer. Fluctuating currents in stochastic thermodynamics II: Energy conversion and nonequilibrium response in kinesin models. *arXiv:1504.03648*, 2015. 26

[75] D. Colquhoun and A. G. Hawkes. Relaxation and fluctuations of membrane currents that flow through drug-operated channels. *Proc. Roy. Soc. Lond. B: Bio. Sci.*, 199(1135):231–262, 1977. 26

[76] S. Lahiri and S. Ganguli. A memory frontier for complex synapses. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Adv. Neural Info. Proc. Sys. 26*, pages 1034–1042. Curran Associates, Inc., 2013. 26

[77] Q. Shen, Q. Hao, and S. M. Gruner. Macromolecular phasing. *Physics Today*, 59(3):46–52, 2006. 26

[78] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. 26

[79] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Correlation-powered information engines and the thermodynamics of self-correction. *Phys. Rev. E*, 95(1):012152, 2017. 26

[80] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Leveraging environmental correlations: The thermodynamics of requisite variety. *J. Stat. Phys.*, in press, arxiv.org:1609.05353, 2016. 26

[81] T. Speck and U. Seifert. The Jarzynski relation, fluctuation theorems, and stochastic thermodynamics for non-Markovian processes. *J. Stat. Mech.: Th. Exp.*, 2007(09):L09002, 2007. 26