Fluctuation theorems from Bayesian retrodiction

Francesco Buscemi [1,*] and Valerio Scarani [2,3]

[1]*Graduate School of Informatics, Nagoya University, Chikusa-ku, Nagoya 464-8601, Japan*
[2]*Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore*
[3]*Department of Physics, National University of Singapore, 2 Science Drive 3, Singapore 117542, Singapore*

Quantitative studies of irreversibility in statistical mechanics often involve the consideration of a reverse process, whose definition has been the object of many discussions, in particular for quantum mechanical systems. Here we show that the reverse channel very naturally arises from Bayesian retrodiction, in both classical and quantum theories. Previous paradigmatic results, such as Jarzynski's equality, Crooks' fluctuation theorem, and Tasaki's two-measurement fluctuation theorem for closed driven quantum systems, are all shown to be consistent with retrodictive arguments. Also, various corrections that were introduced to deal with nonequilibrium steady states or open quantum systems are justified on general grounds as remnants of Bayesian retrodiction. More generally, with the reverse process constructed on consistent logical inference, fluctuation relations acquire a much broader form and scope.

I. INTRODUCTION

In modern statistical mechanics, it has become customary to capture irreversibility by a suitable comparison between a forward and a backward (or reverse) process. Such a comparison is formulated in terms of *fluctuation relations*. Initially limited to linear response, such relations have progressively been extended to encompass a much larger class of processes [1]. After the widely noticed works of Jarzynski [2] and Crooks [3], the literature has grown at such a fast pace that we can only point the reader to some reviews on the matter [4–7].

Forward and backward processes include a prior (initial state), which can be chosen arbitrarily, and a transition rule, *viz.*, a channel. The forward process is the "physical process," i.e., the propagation of the prior through the physical channel: as such, its definition is unproblematic. The main focus of this paper is the construction of the reverse process. For some processes, the identification is clear: for instance, in the case of classical Hamiltonian dynamics, the reverse transition can be easily identified with the dynamics that generates the time-reversed trajectory. For some other processes, however, physical intuition may not be sufficient. Notably, when the transition rule is governed by an underlying quantum channel, there is currently a widespread belief (see, e.g., [7–9]) that Crooks' fluctuation theorem and Jarzynski's equality hold without modifications or other corrections only when the evolution is unitary (i.e., closed) or at least unital (i.e., preserving the uniform distribution). However, nonunital channels such as partial swaps [10–12] or thermal processes [13] are much more obvious paradigms of thermodynamic irreversibility than unital channels. Also, classical fluctuation theorems are

known to hold also for processes that are surely not unital: now quantum theory should be a generalization of classical probability theory, not a restriction over it.

In this paper, we propose that the reverse transition can be systematically constructed as a form of *Bayesian retrodiction* [14–19]. This retrodictive structure can be recognized in some of the most famous classical fluctuations relations [1–3,20,21], even if their original derivation was based on different arguments. Then we prove that, when the forward process is realized as a quantum process (preparation, evolution, and measurement), the Bayesian-reversed process is always realized as a valid quantum process as well. This process coincides with the "quantum retrodiction" independently defined in several works [22–25] and justifies on general grounds the use of Petz's reverse map [26] in the context of fluctuation relations. Our construction applies to any choice of preparation states, quantum channels, and final measurements, leading to exact fluctuation relations that do not require any modification.

The paper is structured as follows. In Sec. II, as preparation, we show how fluctuation relations constitute information-theoretic divergence measures between the forward and the reverse process, and we derive a family of fluctuation relations that originates from Csiszár's f-divergences [27]. In Sec. III we introduce the theory of Bayesian retrodiction and derive the general expression for the classical Bayesian-reversed process. In Sec. IV we consider processes with a quantum realization and show that the Bayesian-reversed process precisely coincides with the quantum retrodiction of the quantum forward process. In Sec. V we recast several known classical fluctuation relations in terms of Bayesian retrodiction. Finally, in Sec. VI we consider the most general case of processes arising from arbitrary quantum channels. In this case we show how our approach very naturally accounts for, and thus justifies as an implicit Bayesian

---

*buscemi@nagoya-u.jp

inversion, all the corrections that were before introduced to deal with situations such as nonequilibrium steady states and nonunital quantum channels.

II. GENERALIZED FLUCTUATION RELATIONS FROM INFORMATION DIVERGENCES

A. Irreversibility as divergence between the forward and the reverse processes

As mentioned, studies of irreversibility involve a comparison between a forward (F) and a reverse (R) process [28]. Consider the evolution of a system from an initial time $t = 0$ to a final time $t = \tau > 0$. Let us assume, for simplicity, that the system possesses a finite state space $\mathcal{A}$. Suppose now that both forward and reverse processes, respectively, are given to us in terms of two suitable joint probability distributions, $P_F(x, y)$ and $P_R(x, y)$, respectively, where $x \in \mathcal{A}$ labels the state of the system at time $t = 0$, while $y \in \mathcal{A}$ labels the state of the system at time $t = \tau$. Concretely, $P_F(x, y)$ denotes the probability that the system starts in state x at time $t = 0$ and ends in state y at time $t = \tau$ under the forward process, while $P_R(x, y)$ denotes the probability that the system starts in state y at time $t = \tau$ and ends in state x at time $t = 0$ under the reverse process. Notice that at this point we are not yet preoccupied with the problem of determining what makes the two processes one the "reverse" of the other. This will be the main issue in the rest of the paper, but for the time being we are just considering $P_F(x, y)$ and $P_R(x, y)$ as given.

If really $P_F(x, y)$ and $P_R(x, y)$ are one the *reverse* of the other, then *irreversibility* may be quantified in terms of how much the two distributions differ, in agreement with the idea that a reversible situation should correspond to the case in which $P_F(x, y) = P_R(x, y)$ for all $x, y \in \mathcal{A}$. In mathematical statistics, the degree of "disagreement" of two distributions is captured in terms of information divergences. Here we focus on the family of f-divergences, defined as[1] [27]

$$D_f(P_F \| P_R) := \sum_{x,y} P_F(x, y) f\left(\frac{P_F(x, y)}{P_R(x, y)}\right)$$
$$= \left\langle f\left(\frac{P_F(x, y)}{P_R(x, y)}\right)\right\rangle_F, \qquad (1)$$

where $f : \mathbb{R}^+ \to \mathbb{R}$ is a function to be specialized further in what follows. For the time being, it suffices to notice that for $f(r) = \ln r$ one recovers the Kullback-Leibler divergence [29], i.e., the usual relative entropy, whereas for $f(r) = r^\alpha$ one recovers the family of Hellinger-Rényi divergences [30]. We will come back to these particular choices at the end of this section.

The forward-reverse ratio will be henceforth denoted

$$r(x, y) = \frac{P_F(x, y)}{P_R(x, y)}. \qquad (2)$$

A priori, one should pay attention to the pairs (x, y) such that this ratio is not well defined. If necessary, this problem can

---

[1]In fact, f-divergences are usually defined in terms of $1/r$; see, e.g., [30]. However, for the sake of the present discussion, formulas are more easily recognizable with the alternative definition.

be dealt with on a case-by-case basis, but there is no need to burden the notation already at this point. Besides, common sense demands (and our later definition will vindicate it) that if a pair (x, y) is assigned a null probability under the forward process, it should be so also under the reverse process, and vice versa. Thus, ultimately, $r(x, y)$ will be ill-defined only for pairs (x, y) that don't contribute to any average because $P_F(x, y) = P_R(x, y) = 0$.

B. f-fluctuation relations

Equation (1) suggests that one interpret $f(r(x, y))$ itself as a *random variable*, whose realization is denoted $\omega_F(x, y)$ for the forward process. The same random variable, when evaluated for the reverse process, is denoted $\omega_R(x, y)$. For consistency, these two must be the *same function*, though evaluated on different arguments: more explicitly,

$$\omega_F(x, y) := f(r(x, y))$$
$$\Downarrow$$
$$\omega_R(x, y) := f\left(\frac{1}{r(x, y)}\right), \qquad (3)$$

simply due to the fact that when considering the reverse variable ω_R, the roles of forward and reverse processes are exchanged and the ratio is thus inverted.

A Crooks-type fluctuation relation is a relation between the probability density functions μ of ω for the forward and the backward processes:

$$\mu_F(\omega) = \sum_{x,y} \delta[\omega_F(x, y) - \omega] P_F(x, y), \qquad (4)$$
$$\mu_R(\omega) = \sum_{x,y} \delta[\omega_R(x, y) - \omega] P_R(x, y)$$
$$= \sum_{x,y} \delta[\omega_R(x, y) - \omega] \frac{1}{r(x, y)} P_F(x, y). \qquad (5)$$

If $f : \mathbb{R}^+ \to \mathbb{R}$ is invertible, everything is well defined, and moreover there exists another function g such that $f(1/r) = g(f(r))$, i.e., $\omega_R(x, y) = g(\omega_F(x, y))$. More explicitly, $g(u) = f(\frac{1}{f^{-1}(u)})$. Plugging these definitions into (5), we obtain

$$\mu_R(u)$$
$$= \sum_{x,y} \delta[\omega_R(x, y) - u] f^{-1}(\omega_R(x, y)) P_F(x, y)$$
$$= f^{-1}(u) \sum_{x,y} \delta\{g[\omega_F(x, y)] - u\} P_F(x, y)$$
$$= \frac{f^{-1}(u)}{|g'(g^{-1}(u))|} \sum_{x,y} \delta[\omega_F(x, y) - g^{-1}(u)] P_F(x, y),$$

and thus, with $u = g(\omega)$, we have a Crooks-type relation of the form

$$\mu_R(g(\omega)) = \frac{f^{-1}(g(\omega))}{|g'(\omega)|} \mu_F(\omega). \qquad (6)$$

Moreover, since $\int_{\mathbb{R}^+} \mu_R(g(\omega)) |g'(\omega)| \, d\omega = \int_{\mathbb{R}^+} \mu_R(u) \, du = 1$, there follows the corresponding Jarzynski-like relation

$$\left\langle f^{-1}(g(\omega)) \right\rangle_F = 1. \qquad (7)$$

The above relation can be verified by unraveling our notations: since ω here is ω_F, we have $g(\omega) = \omega_R$, and $f^{-1}(\omega_R) = 1/r$. Thus $\langle f^{-1}(g(\omega)) \rangle_F = \langle 1/r \rangle_F = \sum_{x,y} \frac{1}{r(x,y)} P_F(x,y) = \sum_{x,y} P_R(x,y) = 1$. Both this direct proof and the derivation from (6) show that, from our perspective, the Jarzynski-like relation (7) expresses the normalization of the reverse process.

C. Examples

The usual choice made in the literature is

$$\omega = f(r) = \frac{1}{z} \ln r,$$

with $z \neq 0$. In this case, we have $f^{-1}(\omega) = e^{z\omega}$, $f(1/r) = -\frac{1}{z} \ln r$, and thus $g(\omega) = -\omega$. The relations (6) and (7) become the familiar ones, namely,

$$\mu_R(-\omega) = e^{-z\omega} \mu_F(\omega) \tag{8}$$

and

$$\langle e^{-z\omega} \rangle_F = 1.$$

Notice that one could have imposed the condition $\omega_R(x,y) = -\omega_F(x,y)$ from the start, in a way somehow reminiscent of an "arrow-of-time" variable. In our notation, this condition is equivalent to impose that the function f satisfies $f(1/r) = -f(r)$, which leads to $f(r) = \frac{1}{z} \ln r$ for an arbitrary z.

As a second example, consider

$$\omega = f(r) = r^\alpha,$$

with $\alpha \neq 0$. Then $f^{-1}(\omega) = \omega^{1/\alpha}$, $f(1/r) = r^{-\alpha}$, and therefore $g(\omega) = 1/\omega$. The fluctuation relations (6) and (7) for random variables that satisfy $\omega_R = 1/\omega_F$ are therefore

$$\mu_R(1/\omega) = \omega^{2-1/\alpha} \mu_F(\omega) \;\Rightarrow\; \langle \omega^{-1/\alpha} \rangle_F = 1.$$

Other examples may look exotic, but nonetheless possible: for instance, by choosing $\omega = f(r) = e^{\kappa r}$, with $\kappa \neq 0$, one obtains $\omega_R = e^{\kappa^2/\ln \omega_F}$, that is, $\ln \omega_R \ln \omega_F = \kappa^2$.

III. REVERSE PROCESS FROM BAYESIAN RETRODICTION

In the previous section we assumed that forward and reverse processes were both given. However, when only the forward process is given, the reverse process should be *derived* from it. In this section we explain how the reverse process can be derived from the forward process by applying the formalism of Bayesian retrodiction.

A. Basics of Bayesian update

We begin this section by reviewing the theory of Bayesian retrodiction. For simplicity, let us consider two random variables X and Y with states labeled by the indices x and y, both taken from a finite set $\mathcal{A}$. Let $P_{XY}(x,y)$ denote their joint distribution. If one's knowledge on Y is updated to a definite value $y = y^*$, the Bayes-Laplace rule tells us that the agents should update their belief on X according to $P'_X(x) = P_{X|Y}(x|y^*)$. In other words, the joint probability is updated as

$$P'_{XY}(x,y) = P_{X|Y}(x|y) \, \delta_{y,y^*}. \tag{9}$$

The above formula, which constitutes the standard formulation of the Bayes-Laplace rule, is silent, however, about those situations, most common in real scenarios, in which the update does not result in a definite value y^*, but is itself described in terms of another probability distribution $P'_Y(y)$. In the face of such "soft evidence," the update should follow *Jeffrey's conditioning* [15]

$$P'_{XY}(x,y) = P_{X|Y}(x|y) P'_Y(y). \tag{10}$$

Jeffrey based this update on his "rule of probability kinematics." The coefficients $P_{X|Y}(x|y)$ are seen as defining a *channel* that propagates the soft belief acquired about y back onto x. It was later noticed that Jeffrey's update can also be obtained from the Bayes-Laplace rule using Pearl's "method of virtual evidence" [16–19]. From this viewpoint, the soft evidence on Y is consequence of another *definite* evidence on a "virtual" variable Z, which has no direct influence on X (i.e., $X \rightarrow Y \rightarrow Z$ forms a Markov chain). When Z is updated to a definite value z^*, one updates the belief on Y to $P'_Y(y) \equiv P_{Y|Z}(y|z) \, \delta_{z,z^*}$ and Eq. (10) is recovered.

B. From update to retrodiction

The Bayesian *update* we just described becomes *retrodiction* when the variables X and Y are given a diachronic meaning: X represents the system's state at the initial time $t = 0$, while Y represents the system's state at the final time $\tau > 0$.

As already noticed, if we know the *forward* transition probability $P_{Y|X}(y|x)$, henceforth denoted as $\varphi(y|x)$, any prior knowledge $p(x)$ on X can be forward-propagated using it. Hence, the *forward process* will be

$$P_F(x,y) = p(x)\varphi(y|x), \tag{11}$$

where $p(x)$ can be chosen arbitrarily. Now we want to define the *reverse* transition probability $\hat{\varphi}(x|y)$, with the same functionality: it should be possible to use it to back-propagate onto X any prior knowledge $q(y)$ that one obtains about Y. In other words, the *reverse process* will be

$$P_R(x,y) = q(y)\hat{\varphi}(x|y), \tag{12}$$

where $q(y)$ can be chosen arbitrarily.

The reverse transition is obtained from Jeffrey's conditioning (10), but the knowledge of the forward transition alone is not enough [31]. One needs to define a *reference prior* $P_X(x)$. In the context of irreversibility, a natural choice is to set the reference prior equal to a steady (*viz.*, invariant) state $\gamma(x)$, that is, a distribution such that

$$\gamma(y) = \sum_x \gamma(x)\varphi(y|x),$$

for all $y \in \mathcal{A}$. This choice coincides with what is customarily done in the theory of Markov chains when defining the reverse chain [32]. Notice that, while every transition matrix possesses at least one steady state, the steady state may not be unique: in such a case, to any choice of a steady state there will correspond a different reverse transition, hence a different fluctuation relation.

Starting from a steady state, the *reverse transition* $\hat{\varphi}(x|y)$ is defined by the relation $\gamma(y)\hat{\varphi}(x|y) = \gamma(x)\varphi(y|x)$. We see that

γ is an invariant distribution for the reverse process too (in other words: by choosing the steady state as prior, we define a reference process that does not distinguish between forward and reverse evolution). Implicitly restricting the analysis to the pairs (x, y) with strictly positive weight [i.e., $\gamma(x)\varphi(y|x) > 0$], we obtain

$$\frac{\hat{\varphi}(x|y)}{\varphi(y|x)} = \frac{\gamma(x)}{\gamma(y)}. \qquad (13)$$

Plugging this together with Eqs. (11) and (12) into (2), we find

$$r(x, y) = \frac{p(x)\gamma(y)}{q(y)\gamma(x)}, \qquad (14)$$

that is, the forward-reverse ratio $r(x, y)$, which is the crucial quantity in the study of irreversibility as noticed in Sec. II, depends on the stochastic transition $\varphi(y|x)$ only through its steady state γ.

IV. QUANTUM INSIDE: RECOVERING QUANTUM RETRODICTION

When the channel is realized by a quantum process, we proceed to show that the formalism of Jeffrey's conditioning is automatically compatible with the formalism of quantum retrodiction [22–25].

Here quantum mechanics enters the picture assuming that the stochastic transition $\varphi(y|x)$ involves an inner quantum "mechanism": to each input $x \in \mathcal{A}$ is associated an input state (density matrix) ρ_0^x, that is later propagated to $\rho_\tau^x = \mathcal{E}(\rho_0^x)$ via a completely positive trace-preserving (CPTP) linear map $\mathcal{E}$, and finally measured using a positive operator-valued measure (POVM) with outcomes $y \in \mathcal{A}$ and elements Π_τ^y. The subscripts 0 and τ are used to denote, respectively, an initial time $t = 0$ and a final time $t = \tau > 0$. With these notations,

$$\varphi(y|x) = \text{Tr}\big[\Pi_\tau^y \, \mathcal{E}\big(\rho_0^x\big)\big]. \qquad (15)$$

In the above equation, the density matrix ρ_0^x is meant to encode all the relevant degrees of freedom needed to represent the xth experimental setup. Hence, it can account for processes in which, for example, there is no clear distinction between system and environment due to the presence of initial correlations [33–38]. In all such cases, ρ_0^x will include not only the degrees of freedom typically associated with the system, but also those associated with the environment.

The expression of $\hat{\varphi}$ in the quantum formalism is immediately obtained from (15). By introducing the state $\gamma_0 = \sum_x \gamma(x)\rho_0^x$, that we assume invertible (otherwise we can restrict the analysis to the subspace where $\gamma_0 > 0$), one gets the much more evocative expression

$$\hat{\varphi}(x|y) = \text{Tr}\big[\Theta_0^x \, \hat{\mathcal{E}}\big(\sigma_\tau^y\big)\big], \qquad (16)$$

which is the Born rule for the POVM elements

$$\Theta_0^x := \gamma(x)\frac{1}{\sqrt{\gamma_0}}\rho_0^x\frac{1}{\sqrt{\gamma_0}}, \qquad (17)$$

the normalized states

$$\sigma_\tau^y := \frac{1}{\gamma(y)}\sqrt{\mathcal{E}(\gamma_0)}\,\Pi_\tau^y\,\sqrt{\mathcal{E}(\gamma_0)}, \qquad (18)$$

and the reverse quantum channel [26,39,40]

$$\hat{\mathcal{E}}(\cdot) := \sqrt{\gamma_0}\,\mathcal{E}^\dagger\bigg[\frac{1}{\sqrt{\mathcal{E}(\gamma_0)}}(\cdot)\frac{1}{\sqrt{\mathcal{E}(\gamma_0)}}\bigg]\sqrt{\gamma_0}, \qquad (19)$$

$\mathcal{E}^\dagger$ being the trace-dual of $\mathcal{E}$, defined by the relation $\text{Tr}[\mathcal{E}^\dagger(X)\,Y] = \text{Tr}[X\,\mathcal{E}(Y)]$ for all operators X and Y. We assume $\mathcal{E}(\gamma_0) > 0$, so that $\hat{\mathcal{E}}$ is a CPTP linear map defined everywhere, and thus physically realizable.[2]

Just as Eq. (13) is the *classical retrodiction* for $\varphi(y|x)$, its quantum description given in Eq. (16) constitutes the *quantum retrodiction* of Eq. (15), in perfect agreement with previous literature on quantum retrodiction [22–25].

V. CLASSICAL THERMODYNAMICS OF RETRODICTION

In this section we show how several important classical fluctuation relations can be recovered using the retrodictive approach.

A. Doubly stochastic transitions and classical Hamiltonian dynamics

We start by noticing that the condition

$$\hat{\varphi}(x|y) = \varphi(y|x) \qquad (20)$$

holds if and only if the transition is doubly stochastic, *viz.*, if in addition to the compulsory normalization $\sum_y \varphi(y|x) = 1$, it also holds $\sum_x \varphi(y|x) = 1$. Indeed, comparing with (13), we see that (20) holds if and only if the steady state can be chosen as the uniform distribution $\gamma(x) \propto 1$. It is then trivial to check that doubly stochastic channels always admit such a steady state, while channels that are not doubly stochastic do not have a uniform steady state.

A particularly important case of doubly stochastic transition is *classical Hamiltonian dynamics*, which is deterministic, *viz.*, there is a one-to-one correspondence between the initial state x and the final state $y \equiv x'$. The evolution $x \to x'$ is then represented by the transition $\varphi(x'|x)$, with $\varphi(x'|x) = 1$ if x' is the final state corresponding to the initial state x, and $\varphi(x'|x) = 0$ otherwise. Thus, for any classical Hamiltonian process one can choose γ as uniform.

For instance, the scenario considered by Bochkov and Kuzovlev [1] is of this type. They specify a class of driven Hamiltonians such that, on any given forward trajectory $x \to x'$, the nondriven term H_0 satisfies $H_0(x') = H_0(x) + E(x, x')$, with E determined by the driving protocol. By choosing thermal priors $p(x) \propto e^{-\beta H_0(x)}$ and $q(x') \propto e^{-\beta H_0(x')}$, with $\beta = 1/k_B T$, they obtain

$$r(x, x') = \frac{\varphi(x'|x)p(x)}{\hat{\varphi}(x|x')q(x')} = \frac{p(x)}{q(x')} = e^{-\beta E(x,x')}. \qquad (21)$$

Since the process is deterministic, the statistics (and in particular the fluctuation relations) carry over from the initial conditions to the whole trajectory. From this observation, rich physical consequences follow: one can derive many-point

---

[2]Even if $\mathcal{E}(\gamma_0)$ is not invertible, $\hat{\mathcal{E}}$ can always be extended to a CPTP map defined everywhere and, thus, physically realizable [41].

relations, relations for the currents (e.g., Onsager), linear response results (in the limit when the driving forces go to zero), etc. [1]. As we presented it, all this can be seen as starting with retrodiction.

Another example that fits in this subsection is that of a generic time-dependent Hamiltonian, but with *microcanonical* priors instead of the most frequently used thermal ones [42]. Staying with a discrete alphabet again (the generalization follows immediately), such priors read: $p(x) = N(E)^{-1}$ if $E_x = E$ (and zero otherwise) and $q(x') = N(E')^{-1}$ if $E_{x'} = E'$ (and zero otherwise), where $N(E)$ and $N(E')$ are the degeneracies of the two levels. Thus, for the processes with nonzero probability, the forward-reverse ratio reads

$$
\begin{aligned}
r(x, x') = \frac{p(x)}{q(x')} &= \frac{N(E')}{N(E)} \\
&= e^{[S(E')-S(E)]/k_B},
\end{aligned} \tag{22}
$$

where $S(E) := k_B \ln N(E)$ coincides with Boltzmann's entropy formula.

B. Classical Hamiltonian system-reservoir interactions

We consider now a system composed of two subsystems. For notational purposes, let us denote the microstates of the first system at initial and final time by the labels x and x', respectively, and those of the second system analogously by the labels w and w'. For the purpose of the present example, all labels belong to discrete sets. If the joint evolution is Hamiltonian, we can borrow from the previous discussion and conclude that

$$
\hat{\varphi}(x, w|x', w') = \varphi(x', w'|x, w). \tag{23}
$$

To construct the forward and reverse processes (11) and (12), one should specify the prior distributions $p(x, w)$ and $q(x', w')$.

This notation opens the possibility of *coarse graining* over one of the subsystems. For definiteness, we describe the particular case studied in [21]. In this narrative, the second system consists of one or several *heat reservoirs*. It is rather natural therefore to stipulate the following:

(1) Both priors are in product form, i.e., $p(x, w) = p(x)P(w)$ and $q(x', w') = q(x')Q(w')$.

(2) The reservoirs' priors $P(w)$ and $Q(w')$ are thermal distributions at inverse temperature $\beta^{-1} = k_B T$, i.e., $P(w) \propto e^{-\beta E_w}$ and $Q(w') \propto e^{-\beta E_{w'}}$, where E_w and $E_{w'}$ are the energies of the reservoir's microstates w and w'. This condition implies

$$
\frac{P(w)}{Q(w')} = e^{\beta(E_{w'}-E_w)} =: e^{\Delta S/k_B} \tag{24}
$$

where ΔS is the entropy generated.

With these two assumptions on the prior distributions, one can compute the following marginal conditional probability:

$$
\varphi(x', \Delta S|x) = \sum_{w,w':E_{w'}-E_w=T\Delta S} \varphi(x', w'|x, w)P(w). \tag{25}
$$

The above represents the forward transition probability that the system, if starting in microstate x, will end up in microstate x' generating in the process an amount of entropy equal to ΔS.

Notice that the coarse-grained transition $\varphi(x', \Delta S|x)$ computed in (25) depends only on the reservoir's prior $P(x)$. In other words, while the reservoir's prior distribution $P(x)$ is fixed, the system's prior remains arbitrary.

Analogously, but starting from the retrodicted transition (23), we obtain

$$
\begin{aligned}
\hat{\varphi}&(x, -\Delta S|x') \\
&= \sum_{w,w':E_w-E_{w'}=-T\Delta S} \hat{\varphi}(x, w|x', w')Q(w') \\
&= \sum_{w,w':E_w-E_{w'}=-T\Delta S} \varphi(x', w'|x, w)Q(w') \qquad (26) \\
&= \sum_{w,w':E_{w'}-E_w=T\Delta S} \varphi(x', w'|x, w)P(w)\frac{Q(w')}{P(w)} \\
&= \sum_{w,w':E_{w'}-E_w=T\Delta S} \varphi(x', w'|x, w)P(w)e^{-\Delta S/k_B} \\
&= e^{-\Delta S/k_B}\varphi(x', \Delta S|x), \qquad (27)
\end{aligned}
$$

where in (26) we used relation (23), while in (27) we used relation (24). The above represents the reverse transition probability that the system, if starting in microstate x', will end up in microstate x generating in the process an amount of entropy equal to $-\Delta S$. Summarizing, we have recovered

$$
\frac{\varphi(x', \Delta S|x)}{\hat{\varphi}(x, -\Delta S|x')} = e^{\Delta S/k_B}, \tag{28}
$$

which is the main result of [21]. It is worth stressing that, as written, ΔS plays the role of an additional random variable coarse graining the reservoir's microstates: that is, $\varphi(x', \Delta S|x)$ describes a transition from input x to output $(x', \Delta S)$. Thus, the object at the denominator in (28) is not the complete Bayesian reverse of the numerator, but rather a *partial* (viz., "hybrid" [21]) reversal.

C. Retrodiction in stochastic thermodynamics

However desirable it would be to derive everything from the Hamiltonian dynamics of a closed (possibly composite) system, information about the latter is often lacking. This is when the approach based on Bayesian retrodiction really shows its power: it allows one to make inferences based on whatever partial information is available.

As a first example of this type, we consider the stochastic thermodynamics setting that led to Crooks' theorem [3]. A general process in discrete-time stochastic thermodynamics is modeled as a sequence of external driving protocols (the *work steps*) alternating with periods during which the system is allowed to equilibrate with an ideal heat bath (the *relaxation steps*) [3,43]. The changes in the system's internal energy during each work step are counted as work done on the system, while the changes happening during each relaxation steps are counted as heat absorbed by the system. For simplicity, we consider here just a two-step process: one work step followed by one relaxation step.

Let us denote the system's initial state by x, and let E_x be the system's initial energy. The system is (deterministically) driven to another energy E'_x. This constitutes the work step.

The relaxation step, which is not deterministic, is modeled using a transition conditional probability $\varphi(y|x)$, where y labels the system's state after the relaxation step. Assuming that during the relaxation step the system's Hamiltonian does not change, the system's final energy, after the relaxation step is over, is E_y'.

By definition, an invariant distribution for $\varphi(y|x)$ is the thermal distribution $\gamma(x) \propto e^{-\beta E_x'}$. This, via (13), leads to the retrodicted transition

$$
\begin{aligned}
\hat{\varphi}(x|y) &= \frac{\gamma(x)}{\gamma(y)} \varphi(y|x) \\
&= e^{\beta(E_y' - E_x')} \varphi(y|x).
\end{aligned}
$$

Choosing as priors the thermal distributions, that is, $p(x) = e^{\beta(F - E_x)}$ and $q(y) = e^{\beta(F' - E_y')}$, with F, F' the corresponding Helmholtz free energies, the ratio (14) becomes

$$
\begin{aligned}
r(x, y) &= \frac{\varphi(y|x) p(x)}{\hat{\varphi}(x|y) q(y)} \\
&= e^{-\beta(E_y' - E_x')} e^{\beta(E_y' - E_x - \Delta F)} \\
&= e^{\beta(E_x' - E_x - \Delta F)} = e^{\beta(W - \Delta F)},
\end{aligned}
$$

where in the last passage we used the assumption that the system's internal energy change during the work step is work W done on the system. The fluctuation relation (8) then immediately gives

$$
\frac{\mu_F(W)}{\mu_R(-W)} = e^{\beta(W - \Delta F)},
$$

in accordance with [3]. Notice that in our retrodictive derivation we did not require any particular condition (such as the condition of detailed balance) for the relaxation step $\varphi(y|x)$, apart from it preserving the thermal distribution, which is implicit in the definition of "relaxation."

VI. QUANTUM THERMODYNAMICS OF RETRODICTION

Our definition of the reverse quantum channel based on Bayesian inversion, presented in Sec. IV, accommodates any state preparation $\{\rho_0^x : x \in \mathcal{A}\}$, any quantum channel $\mathcal{E}$, and any final measurement $\{\Pi_\tau^y : y \in \mathcal{A}\}$. It hence contains, as a very special case, the conventional setup of Tasaki's two-measurement thermodynamics of closed driven quantum systems [44]. But it also contains, and resolves, some quantum setups that have been considered problematic in the past due to the supposed lack of a well-defined reverse process (see, e.g., [7–9]). Where physical intuition fails to envisage a "natural" inversion of the physical process, consistent logical inference (*viz.*, Bayesian inversion) comes to the rescue.

A. Two-measurement setup for closed driven quantum systems

The paradigm for quantum fluctuation relations is provided by Tasaki's *two-measurement setup* [44]. Here a d-level quantum system is prepared in the state ρ_0 but immediately subjected to a von Neumann projective measurement of the initial Hamiltonian $H_0 = \sum_x \epsilon_x |\epsilon_x\rangle\langle\epsilon_x|$. The system, now collapsed onto $|\epsilon_x\rangle$ after observation of ϵ_x, next undergoes a perfectly adiabatic work protocol: its Hamiltonian is driven from H_0 to $H_\tau = \sum_y \eta_y |\eta_y\rangle\langle\eta_y|$, but the system is otherwise perfectly isolated from the surrounding environment. At the

end of the driving protocol, during which only mechanical work has been exchanged with the system, the system is subjected to a second energy measurement, this time of the final Hamiltonian H_τ.

Let us analyze Tasaki's setup with our tools. Denoting the unitary evolution resulting from the driving protocol by $U_{0 \to \tau}$, the forward (predictive) transition probability is given by

$$
\varphi(y|x) = \text{Tr}[U_{0 \to \tau} |\epsilon_x\rangle\langle\epsilon_x| U_{0 \to \tau}^\dagger |\eta_y\rangle\langle\eta_y|]. \quad (29)
$$

It is easy to verify that $\varphi(y|x)$ is doubly stochastic (see Sec. V A). This is a consequence of the following three facts: (1) that the underlying quantum process is unitary, (2) that $\text{Tr}[|\epsilon_x\rangle\langle\epsilon_x|] = \text{Tr}[|\eta_y\rangle\langle\eta_y|] = 1$, and (3) that $\sum_x |\epsilon_x\rangle\langle\epsilon_x| = \sum_y |\eta_y\rangle\langle\eta_y| = \mathbb{1}$. As noted in Sec. V A, the reverse (retrodictive) transition is given by

$$
\varphi(y|x) = \hat{\varphi}(x|y). \quad (30)
$$

At the same time, the corresponding quantum retrodiction, given in Eqs. (17)–(19), is in perfect agreement with a narrative involving time reversals, as in, e.g., [45]: one first prepares the eigenstates of H_τ, then evolves them "backwards in time" using $U_{0 \to \tau}^\dagger$, and finally measures H_0. While such a narrative seems simple and appealing to intuition, it is not directly *operational*, as the actual implementation of time reversals is far from straightforward [46,47]. Moreover, as we will see in what follows, when the evolution between the two energy measurements is not Hamiltonian, a naive argument using time reversal can lead to inconsistencies.

In any case, with Eq. (30) at hand, fluctuation relations can be easily derived. One needs to specify only the two priors $p(x)$ and $q(y)$, which in the two-measurement setup is tantamount to specifying two quantum states ρ_0 and σ_τ, since from them we have $p(x) = \langle\epsilon_x|\rho_0|\epsilon_x\rangle$ and $q(y) = \langle\eta_y|\sigma_\tau|\eta_y\rangle$. Simply by choosing ρ_0 as the thermal state for H_0 and σ_τ as the thermal state for H_τ (by assuming that the temperature is the same), we find that

$$
\begin{aligned}
r(x, y) &= \frac{\varphi(y|x) p(x)}{\hat{\varphi}(x|y) q(y)} \\
&= \frac{p(x)}{q(y)} \\
&= e^{\beta(F_0 - \epsilon_x)} e^{-\beta(F_\tau - \eta_y)} \\
&= e^{\beta(\eta_y - \epsilon_x - \Delta F)} \\
&= e^{\beta(W - \Delta F)},
\end{aligned}
$$

where in the final passage we identified the system's energy difference as work W done on the system, due to the assumption of adiabaticity during the driving protocol. From here, fluctuation relations analogous to Crooks' theorem and Jarzynski's equality quickly follow.

B. Two-measurement setup for general quantum channels and nonequilibrium potentials

Tasaki's setup invites various generalizations. Several references [8,9,48,49] have considered the situation in which the initial and final measurements are as in Tasaki's arrangement but the unitary evolution is replaced by a general CPTP linear map $\mathcal{E}$, leading to the forward transition

$$
\varphi(y|x) = \text{Tr}[\mathcal{E}(|\epsilon_x\rangle\langle\epsilon_x|) |\eta_y\rangle\langle\eta_y|]. \quad (31)
$$

If the linear map $\mathcal{E}$ is not unital, that is, if it does not preserve the unit matrix [*viz.*, $\mathcal{E}(\mathbb{1}) \neq \mathbb{1}$], then the above transition probability is not doubly stochastic in general. For the retrodictive transition, this means that $\hat{\varphi}(x|y) \neq \varphi(y|x)$, and indeed Eq. (13) contains the extra factor $\gamma(x)/\gamma(y)$ coming from the Bayes-Laplace rule.

Nonetheless, a common approach prescribes to stay put with Eq. (30) and work with the *non-normalized* conditional distribution $\tilde{\varphi}(x|y) := \varphi(y|x)$ instead. Obviously, $\tilde{\varphi}(x|y)$ does not admit a realization as in (16), simply because it is not a well-formed conditional distribution. Nonetheless, it is still possible to construct a "ratio"

$$\tilde{r}(x, y) = \frac{\varphi(y|x)p(x)}{\tilde{\varphi}(x|y)q(y)} \equiv \frac{p(x)}{q(y)}$$

and formally go through all the calculations as in the normalized case. However, as a result of considering a reverse "process" that is not properly normalized, the resulting Jarzynski–like relation does not average to 1 as one would like [and as it happens in (7)], but to $\sum_{x,y} \tilde{\varphi}(x|y)q(y)$. This quantity is known as the *efficacy* [8], but we recognize it here as a mathematical artifact arising from an ill-defined reverse transition.

The formalism presented in this work allows a simple treatment of the two-measurement process also for arbitrary quantum channels. The forward transition (31) coincides with Eq. (15) for $\rho_0^x := |\epsilon_x\rangle\langle\epsilon_x|$ and $\Pi_\tau^y := |\eta_y\rangle\langle\eta_y|$. Let $\gamma(x)$ be an invariant distribution for $\varphi(y|x)$. Further, let predictive and retrodictive prior distributions be thermal distributions for some energy levels ϵ_x and η_y, so that $p(x) \propto e^{-\beta\epsilon_x}$ and $q(y) \propto e^{-\beta\eta_y}$. The corresponding initial and final Hamiltonians are defined as $\sum_x \epsilon_x |\epsilon_x\rangle\langle\epsilon_x|$ and $\sum_y \eta_y |\eta_y\rangle\langle\eta_y|$, respectively.

The ratio is defined as usual, that is,

$$\begin{aligned}
r(x, y) &= \frac{\varphi(y|x)p(x)}{\hat{\varphi}(x|y)q(y)} = \frac{p(x)}{q(y)}\frac{\gamma(y)}{\gamma(x)} \\
&= e^{\beta(F_0-\epsilon_x-\ln\gamma(x))}e^{-\beta[F_\tau-\eta_y-\ln\gamma(y)]} \\
&= e^{\beta(\Delta E-\Delta\Phi-\Delta F)},
\end{aligned}$$

where now we find an extra term $\Delta\Phi := \Phi_y - \Phi_x := \frac{1}{\beta}\ln\gamma(x) - \frac{1}{\beta}\ln\gamma(y)$. This term is understood as the difference of a *nonequilibrium potential* that adds to the difference of equilibrium free energy $\Delta F := F_\tau - F_0$. Finally, introducing a total stochastic entropy production $\Sigma := \beta(\Delta E - \Delta\Phi - \Delta F)$, one easily obtains the detailed relation

$$\frac{\mu_F(\Sigma)}{\mu_R(-\Sigma)} = e^\Sigma,$$

and the corresponding integral relation $\langle e^{-\Sigma}\rangle_F = 1$. In this way, we can see how a correct application of the Bayes-Laplace inversion formula automatically takes into account nonequilibrium potentials. These, usually introduced as corrections [20,50], are now recognizable as the remnants of Bayesian inversion.

VII. CONCLUSION

Studies of irreversibility rely on the comparison between a forward physical process and its reverse. We have proposed to define the latter as a form of Bayesian retrodiction (13). We showed that, on the one hand, this definition matches the one used in the derivation of canonical results like Jarzynski's equality, Crooks' theorem, and Tasaki's two-measurement fluctuation theorem for closed driven quantum systems. On the other hand, it also applies to situations in which a reverse process was supposed to be lacking. As a by-product, various modifications, like nonunit efficacies or nonequilibrium potentials, are given a simple explanation as shadows of Bayes-Laplace inversion.

Logical inference thus emerges as a powerful tool to supplement or replace physical intuition, whenever this seems hard to obtain. Clearly, the present approach opens up the possibility for various developments in statistical mechanics and beyond. In particular, an important development is to prove quantum retrodiction as the logical foundation of a fully quantum fluctuation theorem, such as those posited in [51–54], but we leave this for future research.

[1] G. N. Bochkov and Y. E. Kuzovlev, General theory of thermal fluctuations in nonlinear systems, Sov. Phys. JETP **45**, 125 (1977).

[2] C. Jarzynski, Nonequilibrium Equality for Free Energy Differences, Phys. Rev. Lett. **78**, 2690 (1997).

[3] G. E. Crooks, Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems, J. Stat. Phys. **90**, 1481 (1998).

[4] M. Campisi, P. Hänggi, and P. Talkner, Colloquium: Quantum fluctuation relations: Foundations and applications, Rev. Mod. Phys. **83**, 771 (2011).

[5] C. Jarzynski, Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale, Annu. Rev. Condens. Matter Phys. **2**, 329 (2011).

[6] K. Gawedzki, Fluctuation relations in stochastic thermodynamics, arXiv:1308.1518.

[7] K. Funo, M. Ueda, and T. Sagawa, Quantum Fluctuation Theorems, in *Thermodynamics in the Quantum Regime– Fundamental Aspects and New Directions*, edited by F. Binder, L. A. Correa, C. Gogolin, J. Anders, and G. Adesso (Springer International Publishing, Cham, 2018), pp. 249–273.

[8] T. Albash, D. A. Lidar, M. Marvian, and P. Zanardi, Fluctuation theorems for quantum processes, Phys. Rev. E **88**, 032146 (2013).

[9] A. E. Rastegin and K. Zyczkowski, Jarzynski equality for quantum stochastic maps, Phys. Rev. E **89**, 012127 (2014).

[10] S. Lorenzo, R. McCloskey, F. Ciccarello, M. Paternostro, and G. M. Palma, Landauer's Principle in Multipartite Open Quantum System Dynamics, Phys. Rev. Lett. **115**, 120403 (2015).

[11] P. Strasberg, G. Schaller, T. Brandes, and M. Esposito, Quantum and Information Thermodynamics: A Unifying Framework Based on Repeated Interactions, Phys. Rev. X **7**, 021003 (2017).

[12] M. Ziman, P. Stelmachovič, V. Bužek, M. Hillery, V. Scarani, and N. Gisin, Diluting quantum information: An analysis of information transfer in system-reservoir interactions, Phys. Rev. A **65**, 042105 (2002).

[13] M. Horodecki and J. Oppenheim, Fundamental limitations for quantum and nanoscale thermodynamics, Nat. Commun. **4**, 2059 (2013).

[14] S. Watanabe, Symmetry of physical laws. Part III. Prediction and retrodiction, Rev. Mod. Phys. **27**, 179 (1955).

[15] R. C. Jeffrey, *The Logic of Decision* (McGraw-Hill, New York, 1965).

[16] J. Pearls, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Francisco, 1988).

[17] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).

[18] H. Chan and A. Darwiche, On the revision of probabilistic beliefs using uncertain evidence, Artif. Intell. **163**, 67 (2005).

[19] B. Jacobs, The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule, J. Artif. Intell. Res. **65**, 783 (2019).

[20] T. Hatano and S.-I. Sasa, Steady-State Thermodynamics of Langevin Systems, Phys. Rev. Lett. **86**, 3463 (2001).

[21] C. Jarzynski, Hamiltonian derivation of a detailed fluctuation theorem, J. Stat. Phys. **98**, 77 (2000).

[22] S. M. Barnett, D. T. Pegg, and J. Jeffers, Bayes' theorem and quantum retrodiction, J. Mod. Opt. **47**, 1779 (2000).

[23] C. A. Fuchs, Quantum mechanics as quantum information (and only a little more), arXiv:quant-ph/0205039.

[24] M. S. Leifer and Robert W. Spekkens, Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference, Phys. Rev. A **88**, 052130 (2013).

[25] D. Fields, A. Sajia, and J. A. Bergou, Quantum retrodiction made fully symmetric, arXiv:2006.15692.

[26] D. Petz, Sufficiency of channels over von Neumann algebras, Q. J. Math. **39**, 97 (1988).

[27] I. Csiszár, Information type measures of difference of probability distribution and indirect observations, Studia Sci. Math. Hung. **2**, 299 (1967).

[28] U. Seifert, Entropy Production Along a Stochastic Trajectory and an Integral Fluctuation Theorem, Phys. Rev. Lett. **95**, 040602 (2005).

[29] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Statist. **22**, 79 (1951).

[30] F. Liese and K.-J. Miescke, *Statistical Decision Theory* (Springer, New York, 2008).

[31] S. Watanabe, Conditional probabilities in physics, Progr. Theor. Phys. Suppl. **E65**, 135 (1965).

[32] J. R. Norris, *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 1997).

[33] P. Pechukas, Reduced Dynamics Need Not be Completely Positive, Phys. Rev. Lett. **73**, 1060 (1994).

[34] R. Alicki, Comment on Reduced Dynamics Need Not be Completely Positive, Phys. Rev. Lett. **75**, 3020 (1995).

[35] P. Stelmachovic and V. Bužek, Dynamics of open quantum systems initially entangled with environment: Beyond the Kraus representation, Phys. Rev. A **64**, 062106 (2001).

[36] T. F. Jordan, A. Shaji, and E. C. G. Sudarshan, Dynamics of initially entangled open quantum systems, Phys. Rev. A **70**, 052110 (2004).

[37] F. Buscemi, Complete Positivity, Markovianity, and the Quantum Data-Processing Inequality, in the Presence of Initial System-Environment Correlations, Phys. Rev. Lett. **113**, 140502 (2014).

[38] J. M. Dominy, A. Shabani, and D. A. Lidar, A general framework for complete positivity, Quantum Inf. Process. **15**, 465 (2016).

[39] H. Barnum and E. Knill, Reversing quantum dynamics with near-optimal quantum and classical fidelity, J. Math. Phys. **43**, 2097 (2002).

[40] G. E. Crooks, Quantum operation time reversal, Phys. Rev. A **77**, 034101 (2008).

[41] M. M. Wilde, *Quantum Information Theory* (Cambridge University Press, Cambridge, 2013).

[42] P. Talkner, P. Hänggi, and M. Morillo, Microcanonical quantum fluctuation theorems, Phys. Rev. E **77**, 051131 (2008).

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

[44] H. Tasaki, Jarzynski relations for quantum systems and some applications, arXiv:cond-mat/0009244.

[45] T. Sagawa, Second law-like inequalities with quantum relative entropy: An introduction, in *Lectures on Quantum Computing, Thermodynamics and Statistical Physics*, edited by M. Nakahara and S. Tanaka (World Scientific, Singapore, 2012).

[46] A. Gilyén, S. Lloyd, I. Marvian, Y. Quek, and M. M. Wilde, Quantum algorithm for Petz recovery channels and pretty good measurements, arXiv:2006.16924.

[47] M. T. Quintino, Q. Dong, A. Shimbo, A. Soeda, and M. Murao, Reversing Unknown Quantum Transformations: Universal Quantum Circuit for Inverting General Unitary Operations, Phys. Rev. Lett. **123**, 210502 (2019).

[48] J. Goold, M. Paternostro, and K. Modi, Nonequilibrium Quantum Landauer Principle, Phys. Rev. Lett. **114**, 060602 (2015).

[49] A. E. Rastegin, Non-equilibrium equalities with unital quantum channels, J. Stat. Mech. (2013) P06016.

[50] G. Manzano, J. M. Horowitz, and J. M. R. Parrondo, Nonequilibrium potential and fluctuation theorems for quantum maps, Phys. Rev. E **92**, 032129 (2015).

[51] J. Åberg, Fully Quantum Fluctuation Theorems, Phys. Rev. X **8**, 011019 (2018).

[52] Á. M. Alhambra, L. Masanes, J. Oppenheim, and C. Perry, Fluctuating Work: From Quantum Thermodynamical Identities to a Second Law Equality, Phys. Rev. X **6**, 041017 (2016).

[53] H. Kwon and M. S. Kim, Fluctuation Theorems for a Quantum Channel, Phys. Rev. X **9**, 031029 (2019).

[54] G. Manzano, J. M. Horowitz, and J. M. R. Parrondo, Quantum Fluctuation Theorems for Arbitrary Environments: Adiabatic and Nonadiabatic Entropy Production, Phys. Rev. X **8**, 031037 (2018).