Foundations of algorithmic thermodynamics

Aram Ebtekar [*]
Independent Researcher, Vancouver, BC V5Y 3J6, Canada

Marcus Hutter [†]
Google DeepMind, London N1C 4AG, United Kingdom
and School of Computing, Australian National University, Canberra ACT 2601, Australia

(Received 14 December 2023; revised 23 August 2024; accepted 6 December 2024; published 8 January 2025)

Gács's coarse-grained algorithmic entropy leverages universal computation to quantify the information content of any given physical state. Unlike the Boltzmann and Gibbs-Shannon entropies, it requires no prior commitment to macrovariables or probabilistic ensembles, rendering it applicable to settings arbitrarily far from equilibrium. For measure-preserving dynamical systems equipped with a Markovian coarse graining, we prove a number of fluctuation inequalities. These include algorithmic versions of Jarzynski's equality, Landauer's principle, and the second law of thermodynamics. In general, the algorithmic entropy determines a system's actual capacity to do work from an individual state, whereas the Gibbs-Shannon entropy gives only the mean capacity to do work from a state ensemble that is known *a priori*.

DOI: 10.1103/PhysRevE.111.014118

I. INTRODUCTION

Many of the most successful theories in physics have an *initial value formulation*, in which the Universe is fully determined by its *initial conditions* and *dynamical equations of motion*. The second law of thermodynamics, despite being widely considered one of the most important facts of nature, does not appear explicitly in such formulations. Not only is it absent among the dynamical equations, but its irreversibility stands in contrast to the equations' charge-parity-time (CPT) symmetry [1]. Nonetheless, if the second law is to hold for such formulations, then it must somehow follow from the initial condition and dynamics [2–4].

Another difficulty with the second law is its scope of applicability. Informally, it states that the *entropy* of an isolated physical system tends to increase. In order to apply this statement broadly, we require an unambiguous nonequilibrium definition of entropy. Focusing on classical (as opposed to quantum) mechanics for simplicity's sake, our definition should not depend on a prior choice of macrovariables or probabilistic ensemble (as the Boltzmann-Gibbs-Shannon entropies do). Moreover, the second law should not depend on properties such as nonequilibrium steady state or local detailed balance, that only hold in limited settings [5–7].

Thus, we seek to define entropy as a function of the individual states in phase space, in such a way that a suitable initial value formulation makes it increase over time. Our task is made possible by two major insights from the scientific literature: one originating from the theory of dynamical systems, and the other from algorithmic information theory (AIT).

The first insight is the existence of *Markovian coarse grainings*. These are memoryless partitions of a dynamical system's phase space into discrete cells. In more detail, we mean that the probability distribution of the system's coarse-grained state at any future time, conditional on its past and present, is given by evolving the dynamics forward from a uniform (i.e., proportional to the Liouville measure) distribution over the present cell. Whenever this property holds, we can take a discrete view of the system as a time-homogeneous Markov process. Taking the Liouville measure of each cell yields a discrete stationary measure for this process [8].

It remains an open problem to characterize when a coarse graining is approximately Markovian [4,9–13]. To get some intuition, we can study toy systems whose coarse graining is *exactly* Markovian. For instance, Altaner and Vollmer [8] introduce the *network multibaker maps*: these are time-reversible deterministic chaotic dynamical systems, with microscopic randomization only in the initial state, whose coarse graining emulates a wide variety of Markov chains. To illustrate how this occurs, we present a simplified version in Appendix A. These maps rigorously demonstrate that macroscopic irreversibility is compatible with microscopic reversibility, while also hinting at conditions under which more realistic systems might be approximately Markovian.

The Markov assumption is the basis for much of *stochastic thermodynamics* [7,8,14–17], a powerful modern framework that replaces continuous-state dynamical systems by (usually discrete-state) Markov processes. These are easier to analyze: for Markov processes, the nondecrease of Gibbs-Shannon entropy is a straightforward and mathematically rigorous theorem, applicable to probability distributions arbitrarily far from equilibrium ([18], Sec. 4.4).

However, while Markovian coarse grainings motivate a probabilistic description of the *dynamics*, nonequilibrium

[*]Contact author: aramebtech@gmail.com
[†]http://www.hutter1.net/

states may lack an appropriate, nonsubjective ensemble description, especially if they arise from an intricate computation; we elaborate on the reasons in Sec. IV B. To get an ensemble-free definition of entropy, as a function of individual physical states, we turn to a computability-based notion from the AIT literature. Intuitively speaking, the connection between physics and computability arises because the coarse-grained dynamics of our Universe are believed to have computational capabilities equivalent to a universal Turing machine [19–22].

Gács [23] defines the coarse-grained *algorithmic entropy* of any individual state: roughly speaking, it is the number of bits of information that a fixed computer needs in order to identify the state's coarse-grained cell. For example, a state in which all particles are concentrated in one location would have low entropy, because the repeated coordinates can be printed by a short program. If the coarse graining in question is Markovian, then Levin's [24] law of *randomness conservation* says that the algorithmic entropy seldom decreases. In physical terms, we will come to see this as a vast generalization of the second law of thermodynamics.

In mathematical terms, it is an *integral fluctuation* relation, meaning that it bounds an *expectation* of the exponentiated entropy production. It is a statistical law that allows occasional small decreases in entropy. In stochastic thermodynamics, integral fluctuation relations are often derived as averages of corresponding *detailed fluctuation* relations on individual state transitions [14]. In Appendix B we state and prove the detailed fluctuation relations for randomness conservation, and show that they imply the integral relations. The remainder of this article explores the physical consequences of these relations, particularly when dealing with information-processing systems. Thus, we plant the seeds for a new computability-based view of thermodynamics, in which the entropy is a function of individual states rather than probability distributions.

Our article is structured as follows. Section II places our work in the context of some relevant literature. Section III reviews some notation, definitions, and useful facts.

Section IV presents our theoretical contributions. Broadly speaking, we reformulate some aspects of stochastic thermodynamics in terms of Gács's coarse-grained algorithmic entropy. Section IV A adapts Gács's definition to the Markovian setting, where Levin's [24] randomness conservation applies. Section IV B compares the Gibbs-Shannon and algorithmic entropies, presenting conditions under which they coincide. When they do not coincide, we argue that a system's capacity to do work is in fact determined by the algorithmic entropy.

Section IV C introduces the concept of reservoirs that can exchange heat and work, and derives fluctuation inequalities for the algorithmic entropy flow and production. Section IV D specializes these inequalities to the case of one reservoir at constant temperature. One of the inequalities generalizes Kolchinsky's [25] recent lower bound on the heat flow during a computation, while others can be seen as algorithmic versions of Jarzynski's equality [26] and Landauer's principle [27], describing the exchange of heat, work, and information.

Section IV E extends randomness conservation to settings with variable dynamics or long time horizons, resulting in a fully general nonequilibrium second law of thermodynamics. As a special case, we recover a result of Janzing *et al.* [28], but with a different interpretation: we conclude that entropy is only produced when the coarse-grained dynamics are random. Our Corollary 2 is arguably the most complete statement of the second law to date: it uses an ensemble-free notion of entropy, applies to arbitrary time intervals on a trajectory (but not its time-reversal), and allows fluctuations (which include Poincaré recurrence [29]).

Section V presents some more applications. Section V A briefly discusses the effective dynamics of open systems, and introduces a few useful tricks for constructing examples. Section V B applies our algorithmic second law to get an especially straightforward analysis of Maxwell's demon. Section V C discusses the thermodynamic costs of three kinds of information processing: randomization, computation, and measurement. Finally, in order to demonstrate how a deficiency of algorithmic entropy serves as a resource, Sec. V D models an information-theoretic analog of a heat engine, which takes compressible strings as fuel.

Section VI concludes with some possible directions for further research. The core mathematical ideas that we build upon, Markovian coarse grainings and randomness conservation, are detailed in the appendices. While they are based on previous work [8,24,30,31], Appendix A considerably simplifies the network multibaker maps; meanwhile, Appendix B states and proves randomness conservation in a manner that more closely parallels the thermodynamic fluctuation theorems [14]. We hope that these ideas become more accessible as a result.

II. BACKGROUND

When Clausius first coined the term "entropy," thermodynamics was a macroscopic theory of work and heat transfer. The connection to information theory first arose in Szilard's response to Maxwell's famous thought experiment, in which a "demon" appears to violate the second law of thermodynamics by efficiently and intelligently tidying a system. Szilard's great insight was that any such tidying process must necessarily process information about the system. Once this information is correctly accounted for, the second law is restored [32].

This accounting is usually done in an *ad hoc* manner. It remains difficult to define entropy in a unified manner that applies straightforwardly to Maxwell's demon and other nonequilibrium systems. The old definitions of Clausius and Boltzmann depend on specially chosen macrovariables, such as temperature, pressure, and chemical composition; these are well suited to traditional equilibrium systems, but not to the demon's information storage.

Overly fine-grained microscopic definitions of entropy are equally unsuitable. Since the laws of physics are deterministic and time-reversible, fine-grained information can never be created or destroyed. Indeed, in classical mechanics, Liouville's theorem implies that the differential Gibbs-Shannon entropy is constant under Hamiltonian evolutions [33]. While quantum mechanics is beyond the scope of our article, we note that the von Neumann entropy is likewise constant under unitary evolutions. Although Zurek ([34], Appendix C) predicted

a deterministic increase in the algorithmic entropy, which was later proved by Janzing *et al.* [28], Gács [23] shows that this increase is negligible in practice; we will extend these results in Corollary 1.

A more suitable definition lies between the macroscopic and microscopic extremes, as one often finds a *mesoscopic* coarse graining that is approximately Markovian. A standard approach truncates the canonical positions and momenta, thus coarse graining phase space into $6N$-dimensional hypercubes of measure h^{3N}, h being Planck's constant and N the number of particles ([35], Sec. 12 and [17], Sec. 2.7). In this coarse-grained view, the dynamics are effectively random, allowing the entropy to increase at up to the Kolmogorov-Sinai rate [17,36,37]. Nicolis and Nicolis [9] and Werndl [10] study some chaotic systems with Markovian coarse grainings. In quantum mechanical settings, Zurek [38] argues that decoherence has a similar effect.

At present, the Markov assumption appears to be necessary for much of thermodynamics. Indeed, due to CPT symmetry, entropy increase theorems that proceed directly from Hamiltonian dynamics tend to be too weak. Gács [23] discusses a few of these. For example, ergodic systems are known to converge toward maximum entropy in both the infinite future and the infinite past. This fact does not distinguish the two temporal directions and makes no comment on finite time intervals; in particular, it tells us nothing about how the entropy of yesterday should compare to the entropy of tomorrow.

He also discusses another kind of result, in which starting from a "typical" state guarantees that the entropy never falls below its initial value. Kawai *et al.* [39] describe a Hamiltonian formulation (see also Esposito *et al.* [40] for a quantum mechanical version), in which the "typical" state is one where the environment is at equilibrium. For the Universe as a whole, Albert [3] envisions the Big Bang to meet the criteria for a typical state. The problem with this type of argument is that it only works once: subsequent states, having nonminimal entropy, are necessarily "atypical." From there, we have no reason to expect further increases in entropy.

Instead, we need the initial state to have the stronger property that its subsequent transitions forever retain "typical" statistics, long after the state itself becomes atypical. In other words, the coarse-grained trajectory should be a time-homogeneous Markov process. By modeling physical systems as Markov processes, we can formulate the second law over arbitrary time intervals.

From now on, we refer to coarse-grained *states* or *cells* interchangeably. In stochastic thermodynamics, the Gibbs-Shannon entropy is defined as a function, not of individual states, but of *ensembles* that assign a probability $\mu(x)$ to each coarse-grained state x [7,14–17]. Defining the Shannon codelength or *stochastic entropy* (in bits) of each state x by

$$\hat{H}(x,\ \mu) := \log_2 \frac{1}{\mu(x)},$$

the *Gibbs-Shannon entropy* is its expectation

$$H(\mu) := \langle \hat{H}(X,\ \mu)\rangle_{X\sim\mu} = \sum_x \mu(x) \log_2 \frac{1}{\mu(x)}.$$

Note that these entropies are undefined for individual states x, unless a distribution μ is specified. Nonequilibrium choices of μ are typically arrived at by evolving a Markov process starting from a probabilistically prepared initial state. Such approaches have led to a number of major developments in the thermodynamics of information and computation [21,41–46]. Nonetheless, the physical meaning of μ is not always clear [47].

Entropy's main role in physics is to quantify a system's capacity to do work. Inspired by a thought experiment in which compressible data are used to do physical work, Bennett [48] argues that a more precise and general definition of entropy should leverage universal computation to *infer* the best description for any given state. Therefore, he defines the *algorithmic entropy* of an *individual* state x to be its Solomonoff-Kolmogorov-Chaitin description complexity $K(x)$, i.e., the length of the shortest program that outputs x on a fixed universal computer. Li and Vitányi [49] study this function K in detail.

Zurek [34] develops its physical interpretation and proves that

$$H(\mu) \stackrel{+}{=} \langle K(X \mid \mu)\rangle_{X\sim\mu}, \tag{1}$$

where the equality holds up to an additive constant that does not depend on μ. After obtaining a measurement result x from a physical system, Zurek defines its entropy to be the sum $K(x)$ plus the posterior Gibbs-Shannon entropy given x. While this definition appears quite general, its dependence on measurements takes away from the objectivity of Bennett's definition.

Gács [23] proposes a more natural refinement of Bennett's definition. First, he clarifies that the argument x is a coarse-grained state. Second, when states occupy different Liouville measures $\pi(x)$ in the underlying phase space, he adds a correction term, defining the algorithmic entropy as

$$S_\pi(x) := K(x) + \log_2 \pi(x). \tag{2}$$

We can view Gács's definition (2) as a special case of Zurek's, in which the measurement is a fixed function of the fine-grained state. If this function's range is countable, its elements correspond to cells of a coarse graining. Committing to a fixed coarse graining removes the need to actually perform measurements, making (2) purely a function of the individual coarse-grained state x. Thus, measurements are not built into the definition (2), freeing us to consider arbitrary measurements as part of the dynamics, as we illustrate in Sec. V B. We can also consider uncertain or random mesostates, as in stochastic thermodynamics: for random X, the entropy $S_\pi(X) = K(X) + \log_2 \pi(X)$ becomes a random variable that depends on the value of X.

In settings where the Gibbs-Shannon and algorithmic entropies disagree, we must clarify their respective relationships to a system's capacity for work. In Sec. IV D we will see that the algorithmic entropy fundamentally determines the maximum amount of work that can be extracted from a specific physical state. Equation (1) then implies that the Gibbs-Shannon entropy measures the *average* capacity for work, *given* probabilistic knowledge of the state. In Sec. IV B we elaborate on this distinction. Since the algorithmic entropy does not depend on averaging or priors, it carries a more objective physical meaning, which carries over to nonequilibrium settings where we lack probabilistic knowledge.

An important caveat is that the algorithmic entropy depends on a choice of universal computer. The dependence is bounded by the length of a compiler or interpreter between any pair of computers that we want to compare; fortunately, for realistic microprocessors, this length appears to be quite small. Indeed, entropy is measured in logarithmic units, such as bits, nats, or Joules per Kelvin [50]. The conversion rates are given by

$$1\,\text{bit} = k_B \ln 2 = 9.57\times10^{-24}\,\text{J}\,\text{K}^{-1}, \tag{3}$$

where k_B is Boltzmann's constant, equivalent to 1 nat.[1] Pessimistically, consider a pair of computers, whose interpreters in each direction compress to about 12 GiB (i.e., 12×2^{33} bits). This is much larger than any practical interpreter known to the authors; and yet, even languages as distant as these would agree on the entropy of every system to within

$$12\times2^{33}\times9.57\times10^{-24}\,\text{J}\,\text{K}^{-1} < 10^{-12}\,\text{J}\,\text{K}^{-1}.$$

For macroscopic systems, this is a negligible difference. Of course, our notion of a "realistic microprocessor" should include physical size and resource constraints. Inspired by the Turing machine model, we imagine a microscopic control head operating on infinite-length tapes. Since its head is small, such a machine cannot cheat by "hard coding" arbitrary data in its specification. Thus, the laws of nature may well determine which strings are considered simple or complex [19,20]. We hope to make this argument rigorous in future work; see Zurek ([34], Appendix B) for a discussion of related issues.

Our article can also be compared with more recent works. Baez and Stay [53] apply thermodynamics to AIT, whereas we apply AIT to thermodynamics. Kolchinsky and Wolpert [21] relate thermodynamic costs to description complexity; however, their approach depends on specific probabilistic priors, whereas we derive universal bounds from Levin's [24] randomness conservation law. Kolchinsky [25] proves an algorithmic detailed fluctuation inequality, which we show to follow from one of our own.

Throughout this article, we assume the existence of Markovian coarse grainings and a suitable universal computer. The former allows us to substitute physical systems with their Markov process counterparts. Our modeling approach amounts to a minimalist version of stochastic thermodynamics, abstracting away many details of the physics to produce simple rigorous statements under minimal assumptions.

III. PRELIMINARIES

We start with some notation. $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ denote the integers, rational numbers, and real numbers, respectively. $\mathbb{Z}^+$, $\mathbb{Q}^+$, and $\mathbb{R}^+$ denote their respective non-negative subsets. $\mathbb{Z}_m := \{0, 1, \ldots, m-1\}$ denotes the first m elements of $\mathbb{Z}^+$. Let $\mathcal{B} := \{0, 1\} \simeq \mathbb{Z}_2$; its Kleene closure $\mathcal{B}^*$ is the set of all finite-length binary strings. For a string $x \in \mathcal{B}^*$, $|x|$ denotes its length in bits. For a set A, $|A|$ denotes its cardinality. Juxtaposition of strings xy indicates their concatenation. When f is a two-argument function, $f(\cdot, x)$ denotes the one-argument function that maps $y \mapsto f(y, x)$.

The capital letters X, Y refer to random variables, while the lowercase variables x, y refer to the specific values they take on. Expectations of random variables are denoted by angled brackets $\langle\cdot\rangle$. The probability of an event E, conditional on another event F, is denoted by $\Pr(E \mid F)$. The expression $\Pr(Y \mid X)$ is interpreted as a random variable, whose value is $\Pr(Y = y \mid X = x)$ whenever $Y = y$ and $X = x$. Similarly, the conditional expectation $\langle\cdot \mid X\rangle$ is a random variable that depends on the value of X.

A. Stochastic matrices

Stochastic thermodynamics takes place on coarse-grained state spaces, which we represent as countable (i.e., finite or countably infinite) sets $\mathcal{X}, \mathcal{Y}$. Probabilities of transitions from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$ are given by a *stochastic matrix* $P : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^+$, satisfying

$$\forall x \in \mathcal{X}, \quad \sum_{y\in\mathcal{Y}} P(y, x) = 1.$$

Its action on a discrete *measure* $\pi : \mathcal{X} \to \mathbb{R}^+$ takes it to a successor measure $P\pi : \mathcal{Y} \to \mathbb{R}^+$, given by

$$P\pi(y) := \sum_{x\in\mathcal{X}} P(y, x)\pi(x). \tag{4}$$

If $\sum_{x\in\mathcal{X}} \pi(x) = 1$, π is called a *probability measure* (or distribution, mixture, or ensemble), and it follows that $P\pi$ is also a probability measure.

Given π and P such that $P\pi$ is nonzero everywhere, we define the *dual matrix* $\widetilde{P} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ [54,55] by

$$\widetilde{P}(x, y) := \frac{P(y, x)\pi(x)}{P\pi(y)}. \tag{5}$$

Equations (4) and (5) imply that $\widetilde{P}$ is stochastic and satisfies $\widetilde{P}(P\pi) = \pi$.

Note that if x is sampled according to π, and y according to $P(\cdot, x)$, then by Bayes's rule, $\widetilde{P}(x, y)$ gives the reverse transition probability of x given y. However, we will often be interested in nonequilibrium settings, where π is fixed and different from the distribution of x. In that case, the reverse probabilities are sensitive to the distribution of x, and not equal to $\widetilde{P}$ in general.[2]

[1] The 2019 redefinition of SI units made $k_B := 1.380649\times10^{-23}\,\text{J}\,\text{K}^{-1}$ a "defining constant" [51]. Effectively, the Kelvin became a derived unit, equal to *exactly* 1.380649×10^{-23} Joules per nat [52].

[2] On the other hand, P can be viewed as the coarse-grained evolution of an underlying dynamical system, as in Appendix A. Suppose that at a time t_0, the system's state is set to a continuous distribution over its phase space. There is evidence to suggest that the resulting coarse-grained trajectory will evolve forward by P at times $t > t_0 + t_m$, and backward by $\widetilde{P}$ at times $t < t_0 - t_m$. The unpublished manuscript [56] proves this for certain extensions of the multibaker map, establishing a time-reversal symmetry centered near t_0. The "microscopic mixing time" t_m depends on the initial distribution's smoothness, and can be far shorter than the time needed to reach macroscopic equilibrium (i.e., "heat death"). We can think of t_0

For the remainder of this article (except in Appendix B), we take $\mathcal{X} = \mathcal{Y}$. If in addition, $P\pi = \pi$, then we say P is ***π*-stochastic**, or π is *stationary* for P. In that case, $\widetilde{P}$ is also π-stochastic because $\widetilde{P}\pi = \widetilde{P}(P\pi) = \pi$. *Doubly stochastic* is a common synonym for $\sharp$-stochastic, where $\sharp$ denotes the *counting measure*, i.e., $\sharp(x) := 1$ for all x.

Finally, we say that P satisfies *detailed balance* with respect to π if

$$\forall x, y \in \mathcal{X}, \quad P(x,\ y)\pi(y) = P(y,\ x)\pi(x). \tag{6}$$

This is a strong condition, equivalent to having both $P\pi = \pi$ and $P = \widetilde{P}$.

In physical settings, π is the Liouville measure, and $\widetilde{P}$ is the coarse-grained dynamics after CPT inversion. Therefore, detailed balance occurs when all parity-odd microvariables (e.g., momenta) are coarse grained away, leaving only bidirectional transitions [5,7]. In this article, π is stationary by Liouville's theorem, but we allow detailed balance to fail; i.e., $P\pi = \pi$, but possibly $P \neq \widetilde{P}$.

B. Markov processes

An $\mathcal{X}$-valued *stochastic process* is a collection of $\mathcal{X}$-valued random variables indexed by continuous time $(X_t)_{t\in\mathbb{R}^+}$ or by discrete time $(X_t)_{t\in\mathbb{Z}^+}$. We say it is a *time-homogeneous Markov process* if

$$s \leqslant t \Rightarrow \Pr(X_t \mid X_{\leqslant s}) = \Pr(X_t \mid X_s) = P_{t-s}(X_t,\ X_s),$$

where the stochastic matrix $P_{\Delta t} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is called the *transition matrix* for time steps of duration Δt. The first equality is called the *Markov property*, while the second expresses *time homogeneity*. For any finite sequence of times $0 = t_0 < t_1 < \cdots < t_n$, the chain rule of probability yields

$$\Pr\left(X_{t_0},\ \ldots,\ X_{t_n}\right) = \Pr(X_0)\prod_{i=1}^{n} P_{t_i - t_{i-1}}\left(X_{t_i},\ X_{t_{i-1}}\right). \tag{7}$$

A discrete-time Markov process is also called a *Markov chain*. For $\Delta t \in \mathbb{Z}^+$, we have

$$P_{\Delta t} = (P_1)^{\Delta t}. \tag{8}$$

Therefore, (7) implies that a Markov chain's joint probability distribution is uniquely determined by the distribution of X_0 and the matrix P_1; these are its *initial condition* and *dynamics*, respectively. A continuous-time Markov jump process is described similarly, with $P_{\Delta t}$ ($\Delta t \in \mathbb{R}^+$) instead being generated by a transition *rate* matrix [7].

C. Stochastic thermodynamics

We now present a minimalist version of the stochastic thermodynamics framework [7,14–17]. Classical physics studies continuous-time trajectories over the continuous phase space of a dynamical system. The system's evolution is deterministic and reversible, and preserves the phase space's *Liouville measure*. By coarse graining the phase space into discrete cells, we can instead describe the trajectory as a stochastic process (in either discretized or continuous time), that jumps between a countable set of coarse-grained states $\mathcal{X}$. By integrating the Liouville measure over each cell $x \in \mathcal{X}$, we obtain a discrete measure $\pi : \mathcal{X} \to \mathbb{R}^+$.

We assume the coarse-grained trajectory is a Markov process, whose transition probabilities $P_{\Delta t}(y,\ x)$ are equal to the fraction of the cell x, whose evolution after time Δt is in the cell y. This assumption is an intense area of study [8–13], but in stochastic thermodynamics it is commonly taken for granted. It then follows that the Markov process is time-homogeneous and π-stochastic.

To keep things simple and general, we do not make any sort of steady state or detailed balance assumption. At this stage, we do not even define the concepts of energy or heat. In this generic setting, the *stochastic entropy* of a state x, sampled from a probability distribution μ, relative to the (not necessarily normalizable) stationary measure π, is defined by

$$\hat{H}_\pi(x,\ \mu) := \log_2 \frac{\pi(x)}{\mu(x)}. \tag{9}$$

To avoid arithmetic singularities, assume π and μ are nonzero everywhere. We may omit the subscript π in cases where it equals the counting measure $\sharp$. In practice (e.g., see Sec. IV C), the nonuniformity of π comes from modeling environment interactions. Thus, we can think of the contributions $\log_2 \frac{1}{\mu(x)}$ and $\log_2 \pi(x)$ as the entropy (in bits; see (3)) of a base system and its environment, respectively. The generalized *Gibbs-Shannon entropy* of μ is its mean stochastic entropy (i.e., negated Kullback-Leibler divergence) relative to π:

$$H_\pi(\mu) := \langle \hat{H}_\pi(X,\ \mu)\rangle_{X\sim\mu} = \sum_{x\in\mathcal{X}} \mu(x) \log_2 \frac{\pi(x)}{\mu(x)}. \tag{10}$$

The stochastic entropy satisfies the following integral fluctuation theorem.

Theorem 1. Let X, Y be $\mathcal{X}, \mathcal{Y}$-valued random variables, and $P(y,\ x) := \Pr(Y = y \mid X = x)$. If the measures $\pi, \mu : \mathcal{X} \to \mathbb{R}^+$, $\nu : \mathcal{Y} \to \mathbb{R}^+$, and $P\pi$ are nonzero everywhere, then

$$\left\langle 2^{\hat{H}_\pi(X,\ \mu) - \hat{H}_{P\pi}(Y,\ \nu)} \middle| X \right\rangle = \frac{\widetilde{P}\nu(X)}{\mu(X)}.$$

Proof. By definition,

$$\begin{aligned}
\left\langle 2^{\hat{H}_\pi(X,\ \mu) - \hat{H}_{P\pi}(Y,\ \nu)} \middle| X \right\rangle &= \left\langle \frac{\pi(X)\nu(Y)}{P\pi(Y)\mu(X)} \middle| X \right\rangle \\
&= \left\langle \frac{\widetilde{P}(X,\ Y)\nu(Y)}{P(Y,\ X)\mu(X)} \middle| X \right\rangle \\
&= \sum_{y\in\mathcal{Y}} P(y,\ X) \frac{\widetilde{P}(X,\ y)\nu(y)}{P(y,\ X)\mu(X)} \\
&= \frac{\sum_{y\in\mathcal{Y}} \widetilde{P}(X,\ y)\nu(y)}{\mu(X)} \\
&= \frac{\widetilde{P}\nu(X)}{\mu(X)}.
\end{aligned}$$

■

In Theorem 1, μ and ν need not be related to the distribution of (X, Y). Roldán *et al.* [59] apply it to obtain martingales,

as analogous to the Big Bang, and $\widetilde{P}$ as a coarse graining of the conjectured CPT-inverted dynamics preceding it [4,57,58].

essentially by setting $\mu := \widetilde{P}\nu$. However, it is standard to let μ be the initial state distribution. Under the law of a π-stochastic matrix P, the distribution then evolves to $P\mu$. Therefore, the *stochastic entropy production* during a state transition $x \to y$ is defined by

$$\Delta\hat{H}_\pi := \hat{H}_\pi(y,\ P\mu) - \hat{H}_\pi(x,\ \mu).$$

By the law of total expectation, and Theorem 1 with $P\pi = \pi$ and $\nu := P\mu$,

$$\langle 2^{-\Delta\hat{H}_\pi}\rangle = \langle\langle 2^{-\Delta\hat{H}_\pi} \mid X\rangle\rangle = \left\langle \frac{\widetilde{P}P\mu(X)}{\mu(X)} \right\rangle_{X\sim\mu}$$
$$= \sum_{x\in\mathcal{X}} \widetilde{P}P\mu(x) = 1. \tag{11}$$

By Jensen's inequality, it follows that the *Gibbs-Shannon entropy production* is non-negative:

$$\Delta H_\pi := H_\pi(P\mu) - H_\pi(\mu) = \langle\Delta\hat{H}_\pi\rangle \geqslant 0.$$

These results are quite general, applying to distributions that are arbitrarily far from equilibrium. Their main downside is that the stochastic (9) and Gibbs-Shannon (10) entropies depend on a choice of distribution μ and hence are not well-defined functions of the individual state x. While thermodynamic ensembles certainly have their place, Sec. IV B discusses settings in which an adequate choice of μ is not readily available. To get an ensemble-free notion of entropy, we now turn to algorithmic information theory.

D. Algorithmic information theory

A set of strings is *self-delimiting* if no element is a proper prefix of another. For an arbitrary set A, an *encoding* is an injective function $f : A \to \mathcal{B}^*$; we say f is self-delimiting if its range is.

For integers $n \in \mathbb{Z}^+$, a self-delimiting encoding $\overline{n} \in \mathcal{B}^*$ is given recursively by

$$\overline{n} := \begin{cases} \mathtt{0} & \text{if } n = 0, \\ \mathtt{1}\overline{|B(n)|}B(n) & \text{if } n > 0, \end{cases} \tag{12}$$

where $B(n)$ is the standard binary encoding of n without its leading $\mathtt{1}$. For example,

$$\overline{9} = \mathtt{1\bar{3}001} = \mathtt{11\bar{1}1001} = \mathtt{111\bar{0}1001} = \mathtt{11101001}.$$

For strings $x \in \mathcal{B}^*$, a self-delimiting encoding is given by $\overline{x} := \overline{|x|}x$, equal to $\overline{B^{-1}(x)}$ without its leading $\mathtt{1}$. For $A, B \subset \mathcal{B}^*$, let $AB := \{ab : A \in A,\ b \in B\}$. If A is self-delimiting, the pair (a, b) is uniquely decodable from $ab \in AB$. If both A and B are self-delimiting, then so is AB.

A *prefix machine* T is a computer whose set of valid halting programs $\mathcal{P}_T \subset \mathcal{B}^*$ is self-delimiting. Let $T(p)$ denote the output of T on $p \in \mathcal{P}_T$, and write $T(p) = \emptyset$ for $p \in \mathcal{B}^* \setminus \mathcal{P}_T$. We fix a *universal* prefix machine U with the following property: for every prefix machine T, there exists $x_T \in \mathcal{B}^*$, such that for all $y, p \in \mathcal{B}^*$,

$$U(\overline{y}\,\overline{x_T}\,p) = T(\overline{y}p).$$

See Hutter *et al.* [60,61] for the details of this construction, or Li and Vitányi [49] for a similar approach. Since $\overline{y}p$ is uniquely decodable, from now on we instead write (y, p).

The *description complexity* of a string $x \in \mathcal{B}^*$, given side information $y \in \mathcal{B}^*$, is

$$K(x \mid y) := \min_p\{|p| :\ U(y, p) = x\}.$$

When y is the empty string, $K(x \mid y)$ becomes the unconditional description complexity $K(x)$. Whenever an encoding is implied, we may write nonstring objects in place of x or y. For example, a finite set is encoded by a lexicographic listing of its elements. A computable function (or measure or matrix) is encoded by any program that computes it in the sense of (17) below. While the resulting complexity depends on which program is chosen, our derivations remain valid provided that repeated mentions of a function always refer to the same program. We also assume fixed encodings for the countable sets $\mathbb{Z}$, $\mathbb{Q}$, and $\mathcal{X}$, so that their elements have well-defined description complexities. Complexities and entropies in this article are measured in units of bits; accordingly, we use base 2 logarithms.

(In)equalities that hold up to constant additive terms or multiplicative factors are expressed by writing a $+$ or $\times$ on top of the (in)equality sign. For example, $f(x) \overset{+}{<} g(x)$ and $f(x) \overset{\times}{<} g(x)$ mean $f(x) < c + g(x)$ and $f(x) < c \cdot g(x)$, respectively, for some constant c. By "constant," we mean that c is a function of only the parameters that we explicitly declared as fixed, such as the universal computer U and the encodings of important sets. We say x is *simply describable* from an optional context y, if $K(x \mid y) \overset{+}{=} 0$.

We review some properties of K. Since programs are self-delimiting, we have Kraft's inequality

$$\sum_{x\in\mathcal{B}^*} 2^{-K(x|y)} < 1. \tag{13}$$

There exists a decoder p for (12), such that for all $n \in \mathbb{Z}^+$, $U(p\overline{n}) = n$. Hence,

$$K(n) \leqslant |p\overline{n}| \overset{+}{=} |\overline{n}| = 1 + \sum_{i=1}^{\infty} \lfloor 1 + \log_2^i n \rfloor \overset{+}{<} \log_2 n + 2\log_2\log_2 n,$$

where the sum is evaluated as far as the i-fold iterated logarithm is non-negative, and the rightmost inequality assumes $n \geqslant 2$. Similarly, for $x \in \mathcal{B}^*$, $K(x) \overset{+}{<} |\overline{x}| \overset{+}{<} |x| + 2\log_2|x|$.

Whereas the Gibbs-Shannon entropy measures the mean information content per independent sample of a probability distribution [18], the description complexity measures the information content of an individual string without reference to any distribution. Naturally, it satisfies analogous relations [62]. We make frequent use of the following:

$$K(x \mid y) \overset{+}{<} K(x) \overset{+}{<} K(x, y)$$
$$\overset{+}{=} K(y, x) \overset{+}{=} K(y) + K(x \mid y, K(y)) \overset{+}{<} K(y) + K(x \mid y).$$

The algorithmic *mutual information* between x and y is defined by

$$I(x : y) := K(x) + K(y) - K(x, y) \overset{+}{=} K(x) - K(x \mid y, K(y)). \tag{14}$$

The conditional mutual information $I(x : y \mid z)$ is defined by conditioning on z every K term in (14). It can be shown to satisfy a data-processing identity [31]:

$$I(x : (y, z)) \stackrel{+}{=} I(x : z) + I(x : y \mid z, K(z)). \tag{15}$$

Finally, for any computable probability measure μ, and $\delta > 0$,

$$\delta \cdot 2^{-K(x|\mu)} < \mu(x) \stackrel{\times}{<} 2^{-K(x|\mu)},$$

where the first inequality fails on a set of μ probability less than δ,[3] while the second holds for all $x \in \mathcal{B}^*$ ([49], Sec. 4.3). Hence, with μ probability greater than $1 - \delta$,

$$K(x \mid \mu) \stackrel{+}{<} \log_2 \frac{1}{\mu(x)} < K(x \mid \mu) + \log_2 \frac{1}{\delta}. \tag{16}$$

In particular, setting μ to a uniform distribution on any finite set $A \subset \mathcal{B}^*$ reveals that $K(x \mid A) \stackrel{+}{<} \log_2 |A|$, with $K(x \mid A) \stackrel{+}{=} \log_2 |A|$ for all but a constant fraction of $x \in A$. For "simple" μ satisfying $K(\mu) \approx 0$, we have $K(x \mid \mu) \approx K(x)$; hence, (16) lets us interpret $K(x)$ as a universal variant of the Shannon codelength that does not depend on a choice of μ.

Note that we stated all these relations with respect to an arbitrary universal computer U. By applying them to the universal computer $U_z(y, p) := U((z, y), p)$, we see that they remain valid when every mention of K is conditioned on any additional data $z \in \mathcal{B}^*$.

Short programs that output x serve as compressed representations of x. Consider the *universal compression algorithm*: for a fixed time budget, simulate all programs of length up to about $|x| + 2 \log_2 |x|$, and then output the shortest program that halted with output x. As the time budget increases to infinity, the resulting program length decreases to $K(x)$. Thus, we may think of $K(x)$ as the optimum lossless compression achievable for x, in the limit of infinite runtime. In the spirit of Occam's razor, suppose we think of the shortest program as "explaining" x [63]; then we see that explanations are falsified in finite time, but never proven, as we can never rule out the possibility that x's shortest program is among those that have yet to halt.

A function $f : \mathcal{B}^* \to \mathcal{B}^*$ (or between countable sets associated with encodings) is *computable* if there exists a prefix machine T, such that $T(x) = f(x)$ for all $x \in \mathcal{B}^*$. The description complexity K is not computable in this sense; however, our universal compression algorithm is easily adapted to compute a decreasing integer sequence that approaches $K(x)$ from above. Hence, we say that K is *upper semicomputable*.

To extend these concepts to real-valued functions, we say $f : \mathcal{B}^* \to \mathbb{R}$ is *lower (upper) semicomputable* if there exists a computable function $g : \mathcal{B}^* \times \mathbb{Z}^+ \to \mathbb{Q}$, such that $g(x, \cdot)$ is monotonically increasing (decreasing), and

$$\lim_{n \to \infty} g(x, n) = f(x).$$

The real-valued function f is *computable* if it is both lower and upper semicomputable; or equivalently, if there exists a computable function g such that, for any desired level of precision n,

$$|g(x, n) - f(x)| < 2^{-n}. \tag{17}$$

IV. THEORETICAL ANALYSIS

A. Coarse-grained algorithmic entropy

As in Sec. III C, we model a physical system's coarse-grained trajectory by a Markov process on the state space $\mathcal{X}$. We also fix a canonical encoding by which to identify $\mathcal{X}$ with a subset of $\mathcal{B}^*$. It will be convenient to define a conditional version of Gács's [23] coarse-grained algorithmic entropy (2). Therefore, let the *algorithmic entropy* of $x \in \mathcal{X}$, given side information $z \in \mathcal{B}^*$, relative to the stationary measure $\pi : \mathcal{X} \to \mathbb{R}^+$, be

$$S_\pi(x \mid z) := K(x \mid z) + \log_2 \pi(x). \tag{18}$$

Note that (18) is formally identical to the stochastic entropy (9), if we replace the prior $\mu(x)$ with $2^{-K(x|z)}$.

The equilibrium properties of S_π are fairly straightforward. Assuming $Z := \sum_{x \in \mathcal{X}} \pi(x) < \infty$, the normalization $\pi(x)/Z$ is an equilibrium ensemble. Further assuming that $\pi(x)/Z$ is simply describable from z, substituting it into (16) yields

$$K(x \mid z) \stackrel{+}{<} \log_2 \frac{Z}{\pi(x)} < K(x \mid z) + \log_2 \frac{1}{\delta}, \tag{19}$$

where the first inequality holds for every state $x \in \mathcal{X}$, while the second fails with equilibrium probability less than δ. Define the π-randomness deficiency,[4] or *negentropy*, by

$$J_\pi(x \mid z) := \log_2 Z - S_\pi(x \mid z) = \log_2 \frac{Z}{\pi(x)} - K(x \mid z). \tag{20}$$

Then (19) becomes simply

$$0 \stackrel{+}{<} J_\pi(x \mid z) < \log_2 \frac{1}{\delta}. \tag{21}$$

Therefore, the algorithmic entropy S_π has a maximum value of about $\log_2 Z$, and at equilibrium it concentrates near this maximum. Under any of the standard ergodic conditions that make a Markov process converge to π, it follows that the algorithmic entropy tends toward this maximum and then rarely fluctuates away from it.

In the general nonequilibrium setting, the randomness conservation theorem of Levin [24] says that the entropy tends not to decrease. For our purposes, a particular formulation is helpful, which we present as Theorem B 2 in Appendix B. There we consider the transition matrix P that transforms a random earlier state X of the process to a later state Y. If P is π-stochastic, with both π and P being computable, then

[3] This follows by summing over those x which violate the inequality, and applying Kraft's inequality (13).

[4] This concept originates in the algorithmic randomness literature [64–66]. In the language of statistical hypothesis testing, randomness deficiency is the logarithm of an *e-value* or *likelihood ratio* between a null hypothesis $\pi(x)/Z$ and a "universal" alternative hypothesis $2^{-K(x|z)}$ [67–69]. Thus, $J_\pi(x \mid z)$ quantifies how implausible it is that x was sampled from the equilibrium distribution if z was known prior to sampling.

Theorem B 2 says that

$$\left\langle 2^{S_\pi(X|\widetilde{P})-S_\pi(Y|\widetilde{P})}\right\rangle \overset{\times}{<} 1. \tag{22}$$

Moreover, for all $\delta > 0$, with probability greater than $1-\delta$,

$$S_\pi(X \mid \widetilde{P}) - S_\pi(Y \mid \widetilde{P}) \overset{+}{<} \log_2 \frac{1}{\delta}. \tag{23}$$

Two remarks are in order. First, unlike its stochastic analog (11), the algorithmic fluctuation relation (22) need not hold with equality. For example, suppose $P(y,\, x) := 1/|\mathcal{X}|$ for all x, y in some large finite set $\mathcal{X}$. If the earlier state has $K(X \mid \widetilde{P}) \overset{+}{=} 0$ with probability one, then

$$\left\langle 2^{S_\sharp(X|\widetilde{P})-S_\sharp(Y|\widetilde{P})}\right\rangle \overset{\times}{=} \langle 2^{-K(Y|\widetilde{P})}\rangle = \frac{1}{|\mathcal{X}|}\sum_{y\in\mathcal{X}} 2^{-K(y|\widetilde{P})} < \frac{1}{|\mathcal{X}|} \ll 1.$$

Second, (23) bounds the increase in entropy by a constant plus $\log_2(1/\delta)$. The constant term comes from basic properties of K and can be made very small by an appropriate choice of reference computer U ([49], Sec. 3.9). The $\log_2(1/\delta)$ term is also quite small: supposing we tolerate $\delta = 2^{-1000}$, it amounts to a kilobit, which is negligible in terms of physical units (recall (3)). Therefore, at macroscopic scales, we can effectively say that entropy never decreases.

The conditional parameter, consisting of a program for $\widetilde{P}$, remains somewhat of a nuisance. We want a fixed entropy function with which to compare states encountered throughout our system's evolution. Conditioning on $\widetilde{P}$ would be fine if it were solely determined by the laws of physics; but unfortunately, it also depends on the time elapsed between the two snapshots X, Y of our system. In Sec. IV E we derive the appropriate correction for this dependence.

In the meantime, we deal with the more urgent matter of choosing a suitable coarse graining and equipping it with a string encoding. We consider two types of coarse graining: *macroscopic* and *mesoscopic*. Gács [23] focuses primarily on the macroscopic type, letting the elements of $\mathcal{X}$ be the Boltzmann macrostates. In other words, given a physical system, we imagine measuring a few specially designated macrovariables, such as the temperature, pressure, and chemical composition of each of its parts, to a reasonable number of significant figures. The phase space is then partitioned into cells corresponding to every possible joint measurement outcome.

Since each cell $x \in \mathcal{X}$ is determined by the values of its macrovariables, any standard numerical encoding (e.g., binary scientific notation) would work. $K(x \mid \widetilde{P})$ then is quite small: even writing, say, a hundred macrovariables, each to a dozen significant figures, requires only a few kilobits. Therefore, the algorithmic entropy (18) of a *macrostate* x simplifies to

$$S_\sharp(x \mid \widetilde{P}) \approx \log_2 \pi(x),$$

which is just its Boltzmann entropy (in bits).

For physical systems whose Boltzmann macrostates mix sufficiently rapidly, we expect this coarse graining to be approximately Markovian. However, there are systems whose Boltzmann macrostates are not ergodic. For example, computer systems are heavily dependent on the stability of their memory states. As a result, every memory state must be treated separately, even if they occur at the same ambient temperature, pressure, and so on. Here we might use the coarse graining $\mathcal{X} := \mathcal{B}^m$, corresponding to all possible settings of an m-bit memory.

In more general settings, we can obtain a mesoscopic coarse graining by rounding or truncating the values of the canonical microvariables ([35], Sec. 12 and [17], Sec. 2.7). For an N-particle Hamiltonian system, each cell $x \in \mathcal{X}$ is determined by the individual particles' three-dimensional positions and momenta to a high but finite level of precision. In phase space, these cells are tiny $6N$-dimensional hypercubes. The string encoding consists of $6N$ numerical positions and momenta, written in any standard format. The cells have equal Liouville measure, making π constant; by normalizing, we can set $\pi := \sharp$, so that $\log_2 \sharp(x) = 0$ and $\widetilde{P}(x,\, y) = P(y,\, x)$. Therefore, the algorithmic entropy (18) of a *mesostate* x simplifies to

$$S_\sharp(x \mid \widetilde{P}) = K(x \mid \widetilde{P}) \overset{+}{=} K(x \mid P),$$

which is just the description complexity of its canonical microvariables.

We apply the macroscopic coarse graining to *reservoir* systems whose macrostates mix rapidly, and the mesoscopic coarse graining to all other systems. We can think of the mesostates as a refinement of the macrostates: each Boltzmann macrostate corresponds to a large set $B \subset \mathcal{X}$ of mesostates, with Boltzmann entropy $\log_2 |B|$. Since $K(B)$ is small, (16) implies agreement between the algorithmic and Boltzmann entropies: for the vast majority of mesostates $x \in B$,

$$K(x \mid P) \approx K(x \mid P, B) \overset{+}{=} \log_2 |B|. \tag{24}$$

One advantage of the mesoscopic description is that the macrovariables need not be chosen in advance. (24) holds not only for the classical Boltzmann macrostates, but for every simply describable finite set $B \subset \mathcal{X}$ containing x. For example, the entropy of a bookshelf may be estimated by taking B to be the set of configurations compatible with how the books are sorted. For all $x \in B \subset \mathcal{X}$,

$$K(x \mid P) \overset{+}{<} K(B \mid P) + \log_2 |B|, \tag{25}$$

since x is uniquely identified by a description of B, along with a numerical index of size $\log_2 |B|$. In particular, the Boltzmann entropy is only an upper bound on $K(x \mid P)$. The latter may be smaller if the state x has additional structure not captured by the Boltzmann macrovariables.

Generalizing the fixed-size index to a variable-size Shannon code, we also have that for all computable probability measures $\mu : \mathcal{X} \to \mathbb{R}^+$,

$$K(x \mid P) \overset{+}{<} K(\mu \mid P) + \log_2 \frac{1}{\mu(x)}. \tag{26}$$

In algorithmic statistics ([49], Sec. 5.5 and [70–72]), the right-hand sides of (25) and (26) are minimized in order to *infer* which set B or ensemble μ best describes x. There even exist so-called *nonstochastic* strings x, for which no simply describable B or μ makes the inequalities tight [70]. Thus,

the algorithmic entropy takes into account much more general descriptions of states, than do the traditional Boltzmann and Gibbs-Shannon entropies.

B. Comparison against Gibbs-Shannon entropy

One reason why entropy is often regarded as mysterious is that its definitions present conceptual challenges. The Gibbs-Shannon entropy is a function of ensembles and therefore makes no comment on individual physical states. The algorithmic entropy, on the other hand, is defined for all individual states, but fails to be computable. While every program that outputs x yields an *upper* bound on $K(x)$, there is *nox* for which we can compute a large *lower* bound. For if there were, then a small algorithm could search for such x and output the first one it finds, in contradiction to $K(x)$ being large. This is Chaitin's incompleteness theorem ([73], Sec. 4 and [31], Theorem 1.5.2).

While it appears to be a serious shortcoming, the incomputability of K was shown to *benefit* its application to universal prediction ([74], Sec. 3). To see how incomputability likewise benefits thermodynamics, suppose we instead define the entropy as some state function K', for which large lower bounds *are* computable. Then once again, a short program p can search for and output an x for which $K'(x)$ is large. Identifying p with its smallest physical implementation, we expect K' to satisfy the "physical continuity" property $K'(p,\, y) \approx K'(y)$ for all y. Any program that computes x can also be used to erase a preexisting copy of x.[5] As a result, K' would decrease substantially, from $K'(p,\, x) \approx K'(x)$ to $K'(p,\, 0) \approx 0$. Hence, such a K' cannot have a second law.

Having seen why we need K, we are left with the practical matter of estimating it. While lower bounds cannot be computed for any *individual* x, (16) provides a lower bound that holds for *most* samples from any given distribution μ. Moreover, by adding $\langle \log_2 \pi(X) \rangle$ to both sides of (1) and presenting it alongside the definition (10), we obtain

$$H_\pi(\mu) := \langle \hat{H}_\pi(X,\, \mu) \rangle_{X\sim\mu} \stackrel{+}{=} \langle S_\pi(X \mid \mu) \rangle_{X\sim\mu}.$$

Ignoring the middle expression for a moment, this says the Gibbs-Shannon entropy is an *average over* μ, of the algorithmic entropy *conditioned on prior knowledge of* μ. The two stipulations of averaging and prior knowledge cast doubt on whether H_π captures anything objective about physical states. Fortunately, in a wide range of practical settings, there is a natural choice of ensemble $\mu : \mathcal{X} \to \mathbb{R}^+$, corresponding to the individual underlying state $x \in \mathcal{X}$, such that a more direct equivalence holds:

$$H_\pi(\mu) \approx \hat{H}_\pi(x,\, \mu) \approx S_\pi(x). \tag{27}$$

The idea is as follows. We consider μ to be a useful summary of x when three conditions hold:

(1) μ is simply describable, i.e., computable with $K(\mu) \approx 0$.

(2) μ has a notion of *typicality*: most of its samples share some characteristics of interest.

(3) x is among the typical samples with those characteristics.

Condition 1 already implies

$$S_\pi(x \mid \mu) \approx S_\pi(x), \tag{28}$$

removing the need to condition on prior knowledge of μ. It remains only to get rid of the averaging.

Adding $\log_2 \pi(x)$ on all sides of (16) implies that

$$\Pr_{X\sim\mu}(S_\pi(X \mid \mu) \approx \hat{H}_\pi(X,\, \mu)) \approx 1, \tag{29}$$

where we've hidden the tolerances behind approximation ($\approx$) signs for brevity. We take condition 2 (typicality) to mean that the stochastic entropy concentrates near its expectation:

$$\Pr_{X\sim\mu}(\hat{H}_\pi(X,\, \mu) \approx H_\pi(\mu)) \approx 1. \tag{30}$$

In many practical settings, (30) is derived as a consequence of the law of large numbers ([18], Sec. 3).

Condition 3 is that x is typical for μ, which we take to mean that x belongs to the intersection of the high-probability sets given by (29) and (30), that is,

$$H_\pi(\mu) \approx \hat{H}_\pi(x,\, \mu) \approx S_\pi(x \mid \mu). \tag{31}$$

When all three conditions hold, (28) and (31) together imply (27); that is, the Gibbs-Shannon, stochastic, and algorithmic entropies all approximately coincide!

As an example, let μ be the canonical ensemble for a mechanically isolated container of an ideal gas at thermodynamic equilibrium. This ensemble's simple description meets condition 1, while the gas particles' independence and large number ensure condition 2. The overwhelming majority of states encountered at equilibrium are typical in the sense of condition 3. For those states, we conclude that the equivalence (27) holds. The same argument applies to nonequilibrium ensembles, provided that they are simply describable and satisfy the concentration property (30).

On the other hand, we now consider three examples that break each respective condition. We use them to argue that, when the equivalence (27) does not hold, $S_\pi(x)$ is a more physically correct measure of entropy than $H_\pi(\mu)$. For simplicity, let $\pi := \sharp$, so that $S_\pi = K$.

The first violation, where $K(\mu) \gg 0$, is exemplified by letting μ be the point mass on a high-complexity state x. In this case, $K(x) \stackrel{+}{=} I(x : \mu) \stackrel{+}{=} K(\mu) \gg H(\mu) = 0$. Intuitively, the Gibbs-Shannon approach takes μ as an exogenous parameter to specify that we *know* the value of x and can therefore clear it reversibly. In reality, in order to use any sort of knowledge, we must have it physically encoded in a memory device such as our brain. A strength of the algorithmic approach is that it naturally models knowledge as an endogenous part of the physical system. In Sec. V B we see how to model a Maxwell's demon that acquires, and then uses, information about x.

The second violation, where μ lacks a typical set, is exemplified by a robot that flips a hidden coin to decide whether or

[5] Following Bennett [48], the sequence of computations is $(p, x, 0, 0) \to (p, x, x, g) \to (p, 0, x, g) \to (p, 0, 0, 0)$. First, the program is run to produce a second copy of x along with a computation trace g; then the copy of x is used to reversibly erase the original; finally, the program is run backward to clean up its outputs.

not to drain its battery. The probabilistic state μ of the battery and surrounding heat reservoir becomes an equal mixture of the two ensembles corresponding to a full or empty battery. Hence, $H(\mu)$ reaches an intermediate level: a free energy calculation in terms of Gibbs-Shannon entropy would suggest that the robot can do about half a charge worth of work. In reality, the robot will do either zero or a full charge of work. $H(\mu)$ merely predicts the *average* work output among states sampled from μ. In contrast, we see in Sec. IV D that $K(x)$ predicts the work output from a *specific* state x, whether it be zero or a full charge.

The final violation occurs when x is atypical for μ, in such a way that $K(x) \ll \hat{H}(x, \mu)$. Whether this is due to bad inductive priors or because x is nonstochastic [70], the Shannon code for μ is then suboptimal for x. As a result, a Gibbs-Shannon free energy calculation would underestimate the work that a suitable compression algorithm can extract from x. If we take this calculation too seriously, the compression algorithm would appear to violate the second law of thermodynamics.

It is interesting to observe that the so-called *universal measure* $\mathbf{m}(x) := 2^{-K(x)}$ has a "Shannon codelength" of precisely $\hat{H}(x, \mathbf{m}) = K(x)$. This makes it useful as an inductive prior [56,60,61,63,75,76]. However, regardless of whether we normalize $\mathbf{m}$, its Gibbs-Shannon entropy is physically meaningless because it lacks the concentration property (30).[6] For additional comparisons between the probabilistic [18] and algorithmic [49] flavors of information theory, see [62,63,70,77]. We adopt the view of Kolmogorov, who considered algorithmic descriptions to be conceptually prior to (and more general than) probabilistic ones. In his words: "Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character" [78].

C. Interactions with general reservoirs

Now we specialize our framework to the commonly studied setting in thermodynamics, in which a *base system* interacts with an environment composed of one or more *reservoir systems*. Suppose the ith reservoir's macrostate is determined by its *energy* $E_i \in \mathbb{R}$, and possibly some additional macrovariables (e.g., volume and particle count) that we collectively denote by $\mathbf{V}_i \in \mathbb{R}^{d_i}$ ($d_i \in \mathbb{Z}^+$). Let $\pi_i(E_i, \mathbf{V}_i)$ denote the Liouville measure of this macrostate, so that

$$B_i(E_i, \mathbf{V}_i) := k_B \ln \pi_i(E_i, \mathbf{V}_i)$$

is its Boltzmann entropy in physical units.[7] Our coarse-graining formalism restricts the macrovariables to a discrete set of values; nonetheless, if B_i is approximately linear over typical increments in $(E_i, \mathbf{V}_i)$, then we can model it as a differentiable function.

Define the *temperature* T_i by

$$\frac{1}{T_i} := \left(\frac{\partial B_i}{\partial E_i}\right)_{\mathbf{V}_i}. \tag{32}$$

At any state $(E_i, \mathbf{V}_i)$ for which $T_i \neq 0$, the implicit function theorem lets us locally write the energy as a differentiable function $E_i(B_i, \mathbf{V}_i)$, with

$$dE_i = T_i\, dB_i + \left(\frac{\partial E_i}{\partial \mathbf{V}_i}\right)_{B_i} \cdot d\mathbf{V}_i. \tag{33}$$

Thus, the flow of energy into the reservoir is a sum of two contributions. The *heat flow* $đQ_i$ is the energy transferred via microscopic degrees of freedom:

$$đQ_i := T_i\, dB_i = k_B T_i \ln 2\, d \log_2 \pi_i. \tag{34}$$

In our inclusive approach, there is no external driving. Instead, the *work* $đW_i$ is done by the reservoir; it is the energy transferred via changes in the macrovariables $\mathbf{V}_i$:

$$đW_i = -\left(\frac{\partial E_i}{\partial \mathbf{V}_i}\right)_{B_i} \cdot d\mathbf{V}_i. \tag{35}$$

A reservoir whose only macrovariable is energy (i.e., with $d_i = 0$) cannot exchange work and is known as a *heat reservoir*. Conversely, a reservoir whose π_i is a constant function cannot exchange heat, and is known as a *work reservoir*. Substituting (34) and (35) into (33) yields the *first law of thermodynamics*

$$dE_i = đQ_i - đW_i,$$

or after integrating over a given time interval,

$$\Delta E_i = Q_i - W_i. \tag{36}$$

From now on, we use the bold lowercase variable

$$\mathbf{x} := (x, E_1, \mathbf{V}_1, \ldots, E_m, \mathbf{V}_m)$$

to denote the joint coarse-grained state of our base and reservoir systems. It consists of the base system's mesostate $x \in \mathcal{X}$, which we assume to be of unit Liouville measure, along with the macrostates $(E_i, \mathbf{V}_i)$ of m reservoirs. Thus, the mesostates x, the energies E_i, the macrovariables $\mathbf{V}_i$, and the temperatures T_i are all implicitly functions of the joint state $\mathbf{x}$. The mixing within macrostates should be much faster than the transitions between them; this is the main physical assumption which enables us to treat $\mathbf{x}$ as a Markovian state.

Its joint Liouville measure

$$\pi(\mathbf{x}) := \prod_{i=1}^{m} \pi_i(E_i, \mathbf{V}_i) \tag{37}$$

is stationary with respect to the dynamics P. Hence, the joint algorithmic entropy (18) expands to

$$S_\pi(\mathbf{x} \mid \widetilde{P}) := K(\mathbf{x} \mid \widetilde{P}) + \sum_{i=1}^{m} \log_2 \pi_i(E_i, \mathbf{V}_i). \tag{38}$$

[6] Indeed, if $\mathcal{X}$ is simply describable (as a subset of $\mathcal{B}^*$) and large, then $H(\mathbf{m})$ is also large. The program describing $\mathcal{X}$ enumerates its elements, of which the first $x \in \mathcal{X}$ satisfies $\hat{H}(x, \mathbf{m}) = \log_2 \frac{1}{\mathbf{m}(x)} = K(x) \stackrel{+}{=} K(\mathcal{X}) \stackrel{+}{=} 0$. Thus, x has substantial "probability" $\mathbf{m}(x)$, despite having $\hat{H}(x, \mathbf{m}) \ll H(\mathbf{m})$.

[7] For an infinite reservoir, the energy, volume, and Boltzmann entropy would be infinite. Fortunately, we care only about relative changes in these quantities, so we can normalize them to be finite.

Assuming the macrovariables are simply describable, we have

$$K(\mathbf{x} \mid \widetilde{P}) \approx K(x \mid \widetilde{P}) = S_{\sharp}(x \mid \widetilde{P}),$$
$$\log_2 \pi_i(E_i,\ \mathbf{V}_i) \approx S_{\pi_i}(E_i,\ \mathbf{V}_i \mid \widetilde{P}).$$

In other words, the joint algorithmic entropy (38) is approximately the sum of the individual systems' algorithmic entropies. Motivated by these approximations, during a joint state transition $\mathbf{x} \to \mathbf{y}$, we define the base system's *entropy gain* by

$$\Delta K(\mathbf{x} \to \mathbf{y}) := K(\mathbf{y} \mid \widetilde{P}) - K(\mathbf{x} \mid \widetilde{P}).$$

It can be decomposed into a sum

$$\Delta K(\mathbf{x} \to \mathbf{y}) = \Delta_e K(\mathbf{x} \to \mathbf{y}) + \Delta_i K(\mathbf{x} \to \mathbf{y}) \tag{39}$$

of the (reversible) *entropy flow* from the environment

$$\Delta_e K(\mathbf{x} \to \mathbf{y}) := \log_2 \pi(\mathbf{x}) - \log_2 \pi(\mathbf{y}) \tag{40}$$

and the (irreversible) *entropy production*

$$\Delta_i K(\mathbf{x} \to \mathbf{y}) := S_\pi(\mathbf{y} \mid \widetilde{P}) - S_\pi(\mathbf{x} \mid \widetilde{P}). \tag{41}$$

This completes our core definitions. In practice, the stochastic thermodynamics literature seldom mentions the reservoir macrovariables. Instead, it labels the different means by which the base system can transition from x to y, by *mechanisms* i [7], such that the triple $(\mathbf{x},\ y,\ i)$ uniquely determines the resulting joint state $\mathbf{y}$, and $(x,\ y,\ i)$ determines the transition probability

$$P^{(i)}(y,\ x) := P(\mathbf{y},\ \mathbf{x}).$$

Exactly one mechanism occurs per time step, and each has a different outcome, so that

$$\forall \mathbf{x}, \quad \sum_{i,y} P^{(i)}(y,\ x) = \sum_{\mathbf{y}} P(\mathbf{y},\ \mathbf{x}) = 1.$$

For example, in settings where at most one reservoir changes at a time, i would be the index of this reservoir. If the macrovariables of reservoir i are conserved quantities, their changes are equal and opposite to those of the base system during any transition $x \to y$. Now writing

$$\widetilde{P}^{(i)}(x,\ y) := \widetilde{P}(\mathbf{x},\ \mathbf{y}) = \frac{P(\mathbf{y},\ \mathbf{x})\pi(\mathbf{x})}{\pi(\mathbf{y})},$$

the entropy flow (40) can be expressed in terms of just the base system:

$$\Delta_e K(\mathbf{x} \to \mathbf{y}) = \log_2 \frac{\pi(\mathbf{x})}{\pi(\mathbf{y})} = \log_2 \frac{\widetilde{P}(\mathbf{x},\ \mathbf{y})}{P(\mathbf{y},\ \mathbf{x})} = \log_2 \frac{\widetilde{P}^{(i)}(x,\ y)}{P^{(i)}(y,\ x)}. \tag{42}$$

By (39) and (42), the entropy production is given by

$$\begin{aligned}\Delta_i K(\mathbf{x} \to \mathbf{y}) &= K(\mathbf{y} \mid \widetilde{P}) - K(\mathbf{x} \mid \widetilde{P}) + \log_2 \frac{P^{(i)}(y,\ x)}{\widetilde{P}^{(i)}(x,\ y)} \\ &\approx K(y \mid \widetilde{P}) - K(x \mid \widetilde{P}) + \log_2 \frac{P^{(i)}(y,\ x)}{\widetilde{P}^{(i)}(x,\ y)}.\end{aligned} \tag{43}$$

We remark that the detailed balance condition $P = \widetilde{P}$, which expands to (6), is equivalent to having $P^{(i)} = \widetilde{P}^{(i)}$ for all i, which the literature refers to as *local detailed balance* [5–7]. In this article, we do not assume (local) detailed balance. We also opt for the more explicit notation $P(\mathbf{y},\ \mathbf{x})$, in terms of joint states rather than mechanisms.

Having defined our key thermodynamic quantities, we now derive relationships between them, by drawing upon the mathematical results in Appendix B. Theorem B 2 formulates the second law of thermodynamics as an *integral fluctuation* inequality. It says that, regardless of the initial state distribution, the entropy production (41) has a strong statistical tendency to be non-negative:

$$\langle 2^{-\Delta_i K} \rangle \overset{\times}{<} 1. \tag{44}$$

Substituting (40) and (41) into (B6) and (B7) yields a pair of *detailed fluctuation* inequalities. They bound the entropy flow and production for every individual state transition:

$$\Delta_e K(\mathbf{x} \to \mathbf{y}) \overset{+}{<} \log_2 \frac{1}{P(\mathbf{y},\ \mathbf{x})} - K(\mathbf{x} \mid \mathbf{y}, \widetilde{P}), \tag{45}$$

$$-\Delta_i K(\mathbf{x} \to \mathbf{y}) \overset{+}{<} \log_2 \frac{1}{P(\mathbf{y},\ \mathbf{x})} - K(\mathbf{y} \mid \mathbf{x}^*_{\widetilde{P}}, \widetilde{P}). \tag{46}$$

While we obtain $\mathbf{x}^*_{\widetilde{P}} := (\mathbf{x},\ K(\mathbf{x} \mid \widetilde{P}))$ in Appendix B for technical reasons, in practice we can think of $\mathbf{x}^*_{\widetilde{P}}$ as simply $\mathbf{x}$, since their information content is usually about equivalent ([31], Sec. 3.3.2).

To interpret these inequalities, note that the *logical irreversibility* $K(\mathbf{x} \mid \mathbf{y}, \widetilde{P})$ is the amount of information lost about a previous state $\mathbf{x}$, upon transitioning to $\mathbf{y}$. To compensate for the lost information, (45) says that either $\mathbf{y}$ must be a low-probability outcome, or else entropy must flow into the environment. Meanwhile, (46) says that entropy production can only be negative for state transitions that are less likely than their "algorithmic probability" $2^{-K(\mathbf{y} \mid \mathbf{x}^*_{\widetilde{P}}, \widetilde{P})}$.

The inequalities (44) to (46) are information-theoretic in nature, holding for very general environments with arbitrary π. Next, we consider constant temperature environments, for which these inequalities resemble well-known thermodynamic fluctuation theorems.

D. Constant temperature reservoirs

Suppose each reservoir's temperature T_i is constant. Holding $\mathbf{V}_i$ fixed while integrating (33) with respect to B_i yields

$$E_i = T_i B_i + E_{i,\mathrm{work}}, \tag{47}$$

with a "constant of integration" $E_{i,\mathrm{work}}(\mathbf{V}_i)$ that depends only on the macrovariables $\mathbf{V}_i$. The heat and work become exact differentials, since substituting (47) into (35) yields $đW_i = -dE_{i,\mathrm{work}}$, and then integrating yields

$$Q_i = T_i \Delta B_i = k_B T_i \ln 2\, \Delta \log_2 \pi_i = -k_B T_i \ln 2\, \Delta_e K, \tag{48}$$

$$W_i = -\Delta E_{i,\mathrm{work}}. \tag{49}$$

As a result, we no longer need continuous variables to differentiate: provided that E_i changes linearly with B_i, we can *define* the heat and work using Eqs. (47) to (49).

Now, we solve (47) for

$$\ln \pi_i(E_i,\ \mathbf{V}_i) = \frac{B_i(E_i,\ \mathbf{V}_i)}{k_B} = \frac{E_i - E_{i,\mathrm{work}}(\mathbf{V}_i)}{k_B T_i}.$$

Substituting into (37), the joint Liouville measure is

$$\pi(\mathbf{x}) = \prod_{i=1}^{m} \pi_i(E_i,\ \mathbf{V}_i) = \exp\left(\sum_{i=1}^{m} \frac{E_i - E_{i,\text{work}}(\mathbf{V}_i)}{k_B T_i}\right). \quad (50)$$

Up to a normalization factor, (50) corresponds to a number of well-known formulas for the Gibbs measure. For example, consider the case where each reservoir i has not only a constant temperature T_i, but also a constant pressure p_i and chemical potential μ_i for a common species of particle. Then $\mathbf{V}_i$ consists of the reservoir's volume V_i and particle number N_i, and the Gibbs measure is given by (50) with $E_{i,\text{work}}(V_i,\ N_i) := -p_i V_i + \mu_i N_i$ [79]. The mechanical term $-p_i V_i(x)$ and the chemical term $\mu_i N_i(x)$ correspond to the reservoir's ability to do each type of work.

In the case $m = 1$, where there is only one reservoir, the joint state $\mathbf{x}$ is effectively determined as a function of only the base system state x. Formally, let $E_0(x)$, $V_0(x)$, and $N_0(x)$, respectively, denote the energy, volume, and particle number of the base system at state x, and suppose that the totals $E_0(x) + E_1$, $V_0(x) + V_1$, and $N_0(x) + N_1$ are conserved. After normalizing the reservoir macrovariables E_1, V_1, N_1 so that each of the totals is zero,

$$\mathbf{x} = (x,\ E_1,\ V_1,\ N_1) = (x,\ -E_0(x),\ -V_0(x),\ -N_0(x)),$$

Consequently, (50) simplifies to

$$\pi(\mathbf{x}) = \pi(x) = \exp\left(\frac{-E_0(x) - E_{\text{work}}(x)}{k_B T}\right), \quad (51)$$

with $E_{\text{work}}(x) := pV_0(x) - \mu N_0(x)$ in the case of constant pressure p and chemical potential μ.

More generally, we consider any kind of single-reservoir environment at constant temperature T, whose macrovariables are determined as functions of x. (51) still holds, with a possibly different potential function $E_{\text{work}}(x)$. As a result, the joint algorithmic entropy is given by

$$S_\pi(\mathbf{x} \mid \widetilde{P}) \overset{+}{=} S_\pi(x \mid \widetilde{P}) = K(x \mid \widetilde{P}) - \frac{E_0(x) + E_{\text{work}}(x)}{k_B T \ln 2}. \quad (52)$$

It is customary to multiply aggregate entropies by $-k_B T \ln 2$, in order to express them in units of energy. The result is the *total algorithmic free energy*

$$G(x) := E(x) + E_{\text{work}}(x) - K(x \mid \widetilde{P})\, k_B T \ln 2. \quad (53)$$

G serves as a convenient accounting mechanism. Despite being a function of only the base system's state x, it tracks the total entropy production:

$$\Delta G \overset{+}{=} -k_B T \ln 2\ \Delta_i K. \quad (54)$$

As such, G is nonincreasing up to fluctuations. To be precise, (44) and (54) imply

$$\left\langle \exp\left(\frac{\Delta G}{k_B T}\right)\right\rangle = \left\langle 2^{\Delta G / k_B T \ln 2}\right\rangle \overset{\times}{<} 1. \quad (55)$$

It is also useful to consider thermodynamic potentials which track only *some* changes in entropy; we interpret them as *resources* that convert to and from the excluded form(s) of entropy. For example, define the *Helmholtz algorithmic free energy* by

$$F(x) := E(x) - K(x \mid \widetilde{P})\, k_B T \ln 2. \quad (56)$$

During any state transition, (49), (53), and (56) imply

$$\Delta G = \Delta F + \Delta E_{\text{work}} = \Delta F - W. \quad (57)$$

Substituting (57) into the integral fluctuation inequality (55) yields

$$\left\langle \exp\left(\frac{\Delta F - W}{k_B T}\right)\right\rangle \overset{\times}{<} 1. \quad (58)$$

Markov's inequality [or (23)] then implies, with probability greater than $1 - \delta$,

$$\Delta F - W \overset{+}{<} k_B T \ln \frac{1}{\delta}.$$

Since the right-hand side is negligible at macroscopic scales, we see that F measures the base system's capacity for work: free energy must be spent in order for the base system to do work; and conversely, work must be done on the base system in order to replenish free energy.

To get a closer view of the fluctuations in (58), substitute (54) and (57) into the detailed fluctuation inequality (46):

$$\frac{\Delta F - W}{k_B T \ln 2} \overset{+}{<} \log_2 \frac{1}{P(y,\ x)} - K(y \mid x^*_{\widetilde{P}}, \widetilde{P}).$$

Aside from the $\overset{\times}{<}$ sign, (58) is symbolically identical to Jarzynski's [26] equality. However, there are important differences between the two results: our algorithmic free energy F is a trajectory-level quantity, defined as a function of individual states rather than ensembles. Moreover, we take an inclusive view of the work W, as an interaction with the reservoir rather than as external driving.

Finally, to find an explicit connection between information and heat transfer, we change our choice of thermodynamic potential, from F, to the internal entropy K. Using (39) and (48),

$$\Delta_i K = \Delta K - \Delta_e K = \Delta K + \frac{Q}{k_B T \ln 2}.$$

Substituting into the integral fluctuation inequality (44) yields

$$\left\langle 2^{-\left(\Delta K + \frac{Q}{k_B T \ln 2}\right)}\right\rangle \overset{\times}{<} 1. \quad (59)$$

By Markov's inequality [or (23)] again, with probability greater than $1 - \delta$,

$$\Delta K + \frac{Q}{k_B T \ln 2} \overset{+}{>} -\log_2 \frac{1}{\delta}. \quad (60)$$

In order for a digital memory to clear its data, Landauer [27] argued that it must emit at least $k_B T \ln 2$ of heat per erased bit. Since clearing data reduces its description complexity K, we can think of (59) and (60) as mathematically rigorous formulations of Landauer's principle. The impact of heat flow on the energy efficiency of computer hardware depends on the extent to which the flow is reversible; we examine this in Sec. V C.

The fluctuations in (59) are described by the inequality (45), which generalizes the earlier bounds of Zurek [80] and

Kolchinsky [25]. In the case of near-deterministic transitions with negligible complexity [i.e., $\log_2 P(y,\, x) \stackrel{+}{=} K(\widetilde{P}) \stackrel{+}{=} 0$], (45) reduces to Zurek's inequality

$$-\Delta_e K(x \to y) \stackrel{+}{>} K(x \mid y). \tag{61}$$

On the other hand, substituting (48) into (45) yields Kolchinsky's inequality

$$\frac{Q(x \to y)}{k_B T \ln 2} \stackrel{+}{>} K(x \mid y, \widetilde{P}) - \log_2 \frac{1}{P(y,\, x)}. \tag{62}$$

It can be seen as a detailed version of Landauer's principle, giving the minimum heat transfer that accompanies a state transition $x \to y$, in terms of its probability $P(y,\, x)$ and logical irreversibility $K(x \mid y, \widetilde{P})$.

E. Refining the second law

Now, we return to the fully general setting from Sec. IV A, to deal with the algorithmic entropy's dependence on the conditional parameters $\widetilde{P}$. In most applications, the information content of $\widetilde{P}$ consists of a constant part and a variable part. For example, suppose we want to study a particular time-homogeneous π-stochastic Markov chain, at many different times. Then (8) determines the transition matrix $P_{\Delta t}$ in terms of a constant part P_1 and a variable part Δt; to compute its dual $\widetilde{P}_{\Delta t}$, we add π to the constant part.

We regard $(P_1,\, \pi)$ as built into the fundamental laws of physics and assume that they have short encodings on the "natural computers" of the Universe. This assumption can be viewed as a complexity-theoretic version of the physical Church-Turing thesis [21,22], essentially saying that the Universe can implement its own laws on a small computer. Formally, we model this by choosing our reference universal computer in such a way that $K(P_1,\, \pi) \approx 0$. Only the variable part (in this case, Δt) requires an explicit correction.

First, we prove the correction for general $\widetilde{P}$; later, we consider the case where only Δt is variable. To keep this subsection brief, we apply it only to the tail bound (23), though the same correction can also be applied to the integral bound (22) and the detailed bound (46). A reader who is less interested in mathematical details may skip to the main result, Corollary 2.

Theorem 2. Let $\pi : \mathcal{X} \to \mathbb{R}^+ \setminus \{0\}$ be a measure and X, Y be $\mathcal{X}$-valued random variables, such that the matrix $P(y,\, x) := \Pr(Y = y \mid X = x)$ is π-stochastic with a computable dual $\widetilde{P}$. Let $\delta > 0$. Then, with probability greater than $1 - \delta$,

$$\begin{aligned} S_\pi(X) - S_\pi(Y) &\stackrel{+}{<} I(X : \widetilde{P}) - I(Y : \widetilde{P}) + \log_2 \frac{1}{\delta} \\ &\stackrel{+}{<} K(\widetilde{P}) + \log_2 \frac{1}{\delta}. \end{aligned} \tag{63}$$

Proof. For all $x \in \mathcal{X}$, the defining equations (14) and (18) imply

$$\begin{aligned} S_\pi(x \mid \widetilde{P}, K(\widetilde{P})) &= \log_2 \pi(x) + K(x \mid \widetilde{P}, K(\widetilde{P})) \\ &\stackrel{+}{=} \log_2 \pi(x) + K(x) - I(x : \widetilde{P}) \\ &= S_\pi(x) - I(x : \widetilde{P}). \end{aligned} \tag{64}$$

Now, fix $\delta > 0$. We condition Theorem B 2 on the additional data $K(\widetilde{P})$ to find that, with probability greater than $1 - \delta$,

$$S_\pi(X \mid \widetilde{P}, K(\widetilde{P})) - S_\pi(Y \mid \widetilde{P}, K(\widetilde{P})) \stackrel{+}{<} \log_2 \frac{1}{\delta}.$$

In this event, (64) yields

$$\begin{aligned} S_\pi(X) - S_\pi(Y) &\stackrel{+}{=} S_\pi(X_s \mid \widetilde{P}, K(\widetilde{P})) + I(X : \widetilde{P}) \\ &\quad - S_\pi(Y \mid \widetilde{P}, K(\widetilde{P})) - I(Y : \widetilde{P}) \\ &\stackrel{+}{<} I(X : \widetilde{P}) - I(Y : \widetilde{P}) + \log_2 \frac{1}{\delta}. \end{aligned}$$

Now (63) follows from the general bounds $0 \stackrel{+}{<} I(x : z) \stackrel{+}{<} K(z)$. ■

Next, we show that Theorem 2 is tight: in the deterministic and reversible case, its conclusion holds with equality. To be precise, consider the case where P is a permutation matrix; equivalently, the dynamics are described by a bijective transformation $f : \mathcal{X} \to \mathcal{X}$.

Corollary 1. For all computable bijections $f : \mathcal{X} \to \mathcal{X}$, and $x \in \mathcal{X}$,

$$K(f(x)) - K(x) \stackrel{+}{=} I(f(x) : f) - I(x : f).$$

In particular, if $I(x : f) \stackrel{+}{=} 0$, then

$$0 \stackrel{+}{<} K(f(x)) - K(x) \stackrel{+}{<} K(f). \tag{65}$$

Proof. Define a permutation matrix P as follows: for $x, y \in \mathcal{X}$, let $P(y,\, x) := 1$ if $y = f(x)$, and $P(y,\, x) := 0$ otherwise. Since P is doubly stochastic, $\widetilde{P}$ is its transpose, computable by a constant-length program together with f. In order to apply Theorem 2, let $\pi := \sharp$ so that $S_\pi = K$, let the "random" variable X be a constant $x \in \mathcal{X}$ with probability one, and let $Y := f(x)$. Clearly, if the probability of a nonrandom event is positive, then it occurs with certainty. Therefore, by setting $\delta := 1/2$ in Theorem 2,

$$K(x) - K(f(x)) \stackrel{+}{<} I(x : f) - I(f(x) : f) \stackrel{+}{<} K(f).$$

Repeating the same argument for the inverse function yields

$$K(f(x)) - K(x) \stackrel{+}{<} I(f(x) : f) - I(x : f) \stackrel{+}{<} K(f).$$

Combining these inequalities yields the desired conclusions. ■

The leftmost inequality of (65) first appeared in Janzing *et al.* [28]. There the implication

$$I(x : f) \stackrel{+}{=} 0 \Rightarrow K(x) \stackrel{+}{<} K(f(x))$$

was interpreted as saying that the second law of thermodynamics (increase in K) is due to algorithmic independence of the initial condition x from the dynamical law f. The problem with their interpretation lies in the rightmost inequality of (65): f would have to be extraordinarily complex to allow entropy production at a physically meaningful scale. For any deterministic dynamical law that we can feasibly write, Corollary 1 really says that the entropy cannot change by a physically meaningful amount [recall the unit conversions (3)]. Thus, randomness (which in classical physics comes

from coarse graining) is necessary for substantial entropy production to occur.

That being said, we can offer a useful interpretation in line with that of Janzing *et al.* [28]. A doubly stochastic law P can be implemented by choosing a bijection f at random [81], e.g., by repeated tosses of a fair coin. The combination of a low-complexity P, along with the results of sufficiently many coin tosses, then serves as a deterministic law f of very high complexity. Since the coins contribute the bulk of f's information content, algorithmic independence from f really means independence from the coin tosses. If the coins are indeed independent of x, Corollary 1 implies that the entropy cannot decrease, and may increase by up to as many bits as there are coins.

In the setting of a time-homogeneous process, we want to compare the entropies $S_\pi(X_s)$ and $S_\pi(X_t)$, at times $s < t$. The transition matrix P_{t-s} governing the state transition is generated by either P_1 (if time is discrete), or a rate matrix (if time is continuous). A programmatic description of our process's physics should compute the generator and, from it, the transition matrix P_{t-s} and its dual $\widetilde{P}_{t-s}$, over any desired time interval $[s,\, t]$.

Formally, we postulate the existence of a short computer program p, such that for all Δt and $x, y \in \mathcal{X}$, our reference computer U outputs $U(p,\, \Delta t,\, y,\, x) = \widetilde{P}_{\Delta t}(y,\, x)$. In other words, the pair $(p,\, \Delta t)$ is a program for $\widetilde{P}_{\Delta t}$. In the continuous-time case, since the uncountable set $\mathbb{R}^+$ has no encoding, we consider only rational durations $\Delta t \in \mathbb{Q}^+$. By correcting for the description complexity of Δt, we arrive at our most comprehensive, duration-dependent second law of thermodynamics.

Corollary 2 (Algorithmic second law of thermodynamics). Let $\pi : \mathcal{X} \to \mathbb{R}^+ \setminus \{0\}$ be a measure, $(X_t)_{t\in\mathcal{T}}$ be a stochastic process in either continuous $(\mathcal{T} = \mathbb{R}^+)$ or discrete $(\mathcal{T} = \mathbb{Z}^+)$ time, and fix $p \in \mathcal{B}^*$ so that $|p| \stackrel{+}{=} 0$. Consider a pair $s, t \in \mathcal{T}$ with $t - s \in \mathbb{Q}^+$, such that the matrix $P(y,\, x) := \Pr(X_t = y \mid X_s = x)$ is π-stochastic, and its dual satisfies $\widetilde{P}(y,\, x) = U(p,\, t-s,\, y,\, x)$. Then, for $\delta > 0$, with probability greater than $1 - \delta$,

$$\begin{aligned} S_\pi(X_s) - S_\pi(X_t) &\stackrel{+}{<} I(X_s : t-s) - I(X_t : t-s) + \log_2 \frac{1}{\delta} \\ &\stackrel{+}{<} K(t-s) + \log_2 \frac{1}{\delta}. \end{aligned} \tag{66}$$

Proof. Using the pair $(p,\, t-s)$ to encode $\widetilde{P}$, the conclusion of Theorem 2 becomes

$$\begin{aligned} S_\pi(X_s) - S_\pi(X_t) &\stackrel{+}{<} I(X_s : (p,\, t-s)) - I(X_t : (p,\, t-s)) \\ &\quad + \log_2 \frac{1}{\delta} \stackrel{+}{<} K(p,\, t-s) + \log_2 \frac{1}{\delta}. \end{aligned}$$

Since $|p| \stackrel{+}{=} 0$, (66) follows. ∎

We remark that both of the fluctuation terms, $K(t-s)$ and $\log_2(1/\delta)$, are necessary. The former allows periodic visits to low-entropy states, such as in a deterministic process with a very long cycle; whereas the latter allows chance encounters with low-entropy states, such as in a random mixing process. The Poincaré recurrence theorem famously predicts that Hamiltonian systems eventually return to states of low entropy [29]; Corollary 2 is consistent with this finding.[8]

For realistic systems that do not change too quickly, the condition $t - s \in \mathbb{Q}^+$ is not a serious limitation: a small change in t suffices not only to make $t - s$ rational, but also to make its complexity $K(t-s)$ negligible. For concreteness, suppose we restrict our comparisons of entropy to durations that are multiples of Planck's time. In Planck units, since the age of the Universe is less than 2^{203}, all durations up to the present can be represented by integers in the range $1 < \Delta t < 2^{203}$, for which

$$\begin{aligned} K(\Delta t) &\stackrel{+}{<} \log_2 \Delta t + 2 \log_2 \log_2 \Delta t < 203 + 2{\times}8 \text{ bits} \\ &= 219 \text{ bits} < 2.1{\times}10^{-21}\,\mathrm{J\,K^{-1}}. \end{aligned}$$

Therefore, up to negligible fudge terms and exceedingly rare fluctuations, Corollary 2 says that the unconditional entropy S_π is nondecreasing over time.

V. APPLICATIONS

A. Inclusive dynamics and open systems

Our setup thus far is quite general, but cumbersome for the purpose of constructing examples. We focused on inclusive closed system models, in which all influences are explicitly accounted for. As a result, the dynamics are fully determined by the laws of physics, which we assume to be measure-preserving (hence, π-stochastic on a Markovian coarse graining), simply describable, and time-homogeneous. Corollary 2 applies directly to such settings.

On the other hand, sometimes we want to omit the details of some influences, leaving an *open system* model. For example, in the setting of Secs. IV C and IV D, if we model only the base system without the environment's reservoirs, this is an open system. Its mesostates have constant measure, and yet the dynamics have a nonconstant stationary measure π. Since open systems can trade with their environment, measure-preservation and conservation laws do not directly apply.

Without modeling the environment explicitly, can we correct for it to say anything useful? To derive formulas for the entropy flow and production, we must start from the corresponding inclusive model and then eliminate the environment measure π. We did this in (42) and (43); these formulas still depend on π through the dual matrix, but that too is eliminated

[8] That said, entropy fluctuations at the scale of Poincaré recurrence are incredibly rare. To illustrate, consider $10^{10^{20}}$ consecutive time steps of any simply describable duration, be they seconds or years. If we set $\delta := (10^{10^{20}})^{-3}$ in Corollary 2, then by a union bound over all pairs (s, t) of these times, the probability of (66) failing for even *one* pair is less than one in $10^{10^{20}}$. Since $10^{10^{20}} < 2^{3.33\times10^{20}}$, (66) bounds the largest entropy decrease by

$$\begin{aligned} K(t-s) + \log_2 \frac{1}{\delta} &< 3.33{\times}10^{20} + 2\log_2(3.33{\times}10^{20}) \\ &\quad + 9.99{\times}10^{20} < 1.333{\times}10^{21} \text{ bits} < 0.013\,\mathrm{J\,K^{-1}}. \end{aligned}$$

Each increment of the topmost exponent in $10^{10^{20}}$ would only multiply this bound tenfold.

in the case of (local) detailed balance. We can then apply results such as Theorem 2 and Corollary 2, by substituting the correct formula (43) for entropy production.

Other open systems may experience algorithmically complex or time-dependent dynamics. Again, we begin with the inclusive model, and then try to eliminate the influence. Consider a *control system* with state space $\mathcal{C}$, that influences a *base system* with state space $\mathcal{X}$. Suppose the control state stays still, i.e., the joint transition matrix $P : (\mathcal{C}\times\mathcal{X})\times(\mathcal{C}\times\mathcal{X}) \to \mathbb{R}^+$ satisfies

$$c \neq d \Rightarrow P((d, y),\ (c, x)) = 0.$$

For any coarse-grained Liouville measure

$$\pi(c, x) := \pi_{\mathcal{C}}(c)\pi_{\mathcal{X}}(x) \quad \text{for } c \in \mathcal{C},\ x \in \mathcal{X},$$

it is easy to verify that P is π-stochastic iff each of the submatrices $P((c, \cdot),\ (c, \cdot))$ are $\pi_{\mathcal{X}}$-stochastic.

In this manner, a single *global dynamics* P can emulate a wide variety of *local dynamics* $P((c, \cdot),\ (c, \cdot))$ on our base system. Since the submatrices are $\pi_{\mathcal{X}}$-stochastic, results such as Theorem 2 and Corollary 2 apply to our base system. However, the submatrix is not fully determined by the laws of physics P; it also depends on the control input c. If $K(c)$ is sufficiently small, it is safe to ignore this dependence.

However, if $K(c)$ is large, the control might represent a complex piece of information, such as the mind of a Maxwell's demon. In the next subsection, we revisit the famous thought experiment using our algorithmic inequalities. If we insist on modeling x as an open system, then the algorithmic entropy must be conditioned on c.

A useful variant of a control is a reversible program counter or *clock* $\mathcal{C} := \mathbb{Z}_m$, with $\pi_{\mathcal{C}} := \sharp$. Instead of staying still, it ticks ahead in a predetermined manner:

$$c + 1 \not\equiv d \pmod m \Rightarrow P((d, y),\ (c, x)) = 0.$$

Although the global dynamics P is time-homogeneous, a clock enables the base system to undergo a sequence of different $\pi_{\mathcal{X}}$-stochastic transitions, given by the submatrices $P((c + 1, \cdot),\ (c, \cdot))$. If m is not too large, then $K(c) \overset{+}{<} 2\log_2 m$ is quite small.

We can apply Theorem 2 to systems with time-dependent dynamics. Consider a time interval $[s, t]$, during which the evolution is given by a short sequence of simply describable time-dependent transition matrices. Then, setting $(X, Y) := (X_s, X_t)$ in Theorem 2, we have $K(\widetilde{P}) \overset{+}{=} 0$. Allowing for a small but constant probability δ of failure, its conclusion simplifies to $S_{\pi_{\mathcal{X}}}(X_s) \overset{+}{<} S_{\pi_{\mathcal{X}}}(X_t)$. Recall from Sec. IV A that, for nonreservoir systems, we coarse grain into equal-sized mesostates. Therefore, in the absence of heat transfer, the dynamics are doubly stochastic, the algorithmic entropy is $S_\sharp = K$, and the theorem's conclusion becomes

$$K(X_s) \overset{+}{<} K(X_t). \tag{67}$$

It may also seem cumbersome to construct interesting examples of transition matrices. Fortunately, by Révész's generalization of the Birkhoff–von Neumann theorem, transitioning by a doubly stochastic matrix is equivalent to transitioning by a probabilistic mixture of deterministic bijections [81]. This means, instead of writing an explicit transition matrix P, we can describe the dynamics as a *random bijection* $x \mapsto F(x)$, where the bijection $F : \mathcal{X} \to \mathcal{X}$ is sampled independently at each time step, from some distribution with low description complexity.[9] If we only care to specify F on a proper subset of $\mathcal{X}$, then it can be a random *injection* whose domain and range have complements of equal cardinality, since these can be extended to bijections on all of $\mathcal{X}$.

For example, on the two-element state space $\mathcal{X} := \mathcal{B} = \{0, 1\}$, there are exactly two bijections: identity and negation. Therefore, these are the only deterministic dynamics permitted on $\mathcal{B}$. The full set of permitted dynamics (in the absence of heat transfer) are the mixtures of identity and negation, parametrized by a probability of negation $\alpha \in [0, 1]$. Injections on proper subsets of $\mathcal{X}$ are also allowed: for example, we can specify that `0` maps to `1`, without caring what `1` maps to (though in this case, negation is the only bijective extension).

In summary, the behavior of open systems can be influenced in a variety of ways. Complex controls require careful accounting, but simple controls and clocks just extend our modeling capabilities: they allow us to consider systems that evolve by sequences of deterministic or random bijections on $\mathcal{X}$, or injections on subsets of $\mathcal{X}$. The transition function at each time step is sampled from a simply describable distribution. If not too many time steps are taken, then (67) holds with a high probability. This time-dependent setting is flexible enough to illustrate a number of phenomena regarding the thermodynamics of information.

B. Maxwell's demon

As a warmup to more difficult examples, we now review the original challenge to the second law [32]. The core ideas here are not new, but we hope that our simple abstract presentation lends some pedagogical clarity.

In the famous thought experiment, *Maxwell's demon* has a memory that starts in a low-entropy "clear" state $0 \in \mathcal{C}$. It interacts with a base system that starts in some high-entropy mesostate $x \in \mathcal{X}$. It is helpful to begin with a stylized special case, in which $\mathcal{C} = \mathcal{X}$ and the demon is able to reversibly perform a complete measurement, copying the system's state into memory:

$$(0,\ x) \mapsto (x,\ x).$$

[9] Strictly speaking, we should write $F : \Omega \times \mathcal{X} \to \mathcal{X}$ to express dependence on an ambient probability space Ω. If $\mathcal{X}$ is countably infinite, then there are uncountably many bijections on it, so the probability measure need not be discrete. Instead, it can be represented by a uniformly computable sequence of functions $\Gamma_n : \mathcal{X}^{\mathcal{X}_n} \to \mathbb{R}$, where $\mathcal{X}_n \subset \mathcal{X}$ consists of the states whose encodings have length at most n, and $\mathcal{X}^{\mathcal{X}_n}$ is the countable set of functions $\tilde{f} : \mathcal{X}_n \to \mathcal{X}$. We take $\Gamma_n(\tilde{f})$ to be the probability of sampling a function $f : \mathcal{X} \to \mathcal{X}$ whose restriction to $\mathcal{X}_n$ is $\tilde{f}$. In order to compute the transition matrix entry $P(y, x)$ from Γ, let $n = |x|$, and enumerate functions $\tilde{f} : \mathcal{X}_n \to \mathcal{X}$ until their total probability is as close to 1 as desired. Then $P(y, x)$ is approximately the probability assigned to functions that satisfy $\tilde{f}(x) = y$. Since $K(P) \overset{+}{<} K(\Gamma)$, if Γ is simply describable, then so is P.

Using its measurement as a control, the demon proceeds to reversibly erase the base system's entropy:

$$(x,\ x) \mapsto (x,\ 0).$$

Both of these mappings are injective on their respective domains, $\{(0,\ x) : x \in \mathcal{X}\}$ and $\{(x,\ x) : x \in \mathcal{X}\}$. Therefore, they can be extended to bijections on $\mathcal{X} \times \mathcal{X}$. Since they are deterministic, Corollary 1 implies the total entropy cannot change substantially. Indeed,

$$K(0,\ x) \stackrel{+}{=} K(x,\ x) \stackrel{+}{=} K(x,\ 0) \stackrel{+}{=} K(x).$$

The net effect is that the base system's information content is moved into the demon's memory. The erasure step uses a high-entropy control x; as discussed in Sec. V A, its entropy must be included in the total, for otherwise the base system's transition $x \mapsto 0$ would appear to violate the second law. The erasure is permitted precisely because it occurs in the presence of a copy. The second law forbids the demon from clearing the last copy remaining in its memory:

$$(x,\ 0) \mapsto (0,\ 0),$$

as can also be seen by noting that this map is either not injective (if defined to work for all x), or not simply describable (if tailored for a specific x).

Now we present the general case, which includes nearly all physical models of Maxwell's demon from the literature. Suppose the demon performs a partial measurement $m(x)$, where m is a (possibly random, not necessarily bijective) function, whose distribution has low description complexity. After obtaining the measurement, the demon uses it as a control to transition the system from x to some new (possibly random) state y:

$$(0,\ x) \mapsto (m(x),\ x) \mapsto (m(x),\ y).$$

The specifics of y's computation are not important: as long as it amounts to a simply describable mixture of bijections, the second law expressed by (67) holds with high probability. It expands to

$$K(m(x),\ x) \stackrel{+}{<} K(m(x),\ y).$$

Subtracting $K(m(x))$ from both sides yields

$$\begin{aligned} K(x) - I(m(x) : x) &\stackrel{+}{=} K(x \mid m(x)^*) \stackrel{+}{<} K(y \mid m(x)^*) \\ &\stackrel{+}{=} K(y) - I(m(x) : y). \end{aligned} \tag{68}$$

We can interpret the conditional complexities as subjective entropies, from the point of view of a demon that knows the measurement $m(x)$. The measurement makes the base system's subjective entropy less than its objective (i.e., unconditional) entropy.

Now, (68) implies $K(x) - K(y) \stackrel{+}{<} I(m(x) : x)$. Consider the case where the mutual information of measurement is reversibly erased: "reversibly" meaning $K(m(x),\ x) \stackrel{+}{=} K(m(x),\ y)$, and "erased" meaning $I(m(x) : y) \stackrel{+}{=} 0$. Then, in fact,

$$K(x) - K(y) \stackrel{+}{=} I(m(x) : x).$$

Thus, although the second law forbids a decrease in the *total* entropy, it permits the measured system to lose as much entropy as was measured from it! There is no contradiction here: since the algorithmic entropy is subadditive, it is possible to have simultaneously $K(x) \gg K(y)$ and $K(m(x),\ x) \stackrel{+}{<} K(m(x),\ y)$.

Note that our analysis does not require the completion of a cycle, nor any ad hoc extension of the definition of entropy. The algorithmic entropy naturally accounts for both the demon's memory and the base system, satisfying the second law of thermodynamics at every step of their evolution.

C. Landauer's principle

In Sec. IV D we saw that a system can dump its unwanted entropy into a reservoir as heat. Landauer [27] first discovered this in the context of computer circuitry, arguing that logically irreversible computations necessarily convert some energy into waste heat. Neyman [82] followed up with a larger bound in settings involving irreversible thermal equilibration. There are now many modern references treating Landauer's principle [43,83,84].

We can develop similar ideas in terms of the algorithmic entropy. Just as (39) decomposes a base system's entropy gain into reversible and irreversible parts, we can rearrange (39) to decompose the *environment*'s entropy gain into reversible and irreversible parts:

$$-\Delta_e K(\mathbf{x} \to \mathbf{y}) = \Delta_i K(\mathbf{x} \to \mathbf{y}) - \Delta K(\mathbf{x} \to \mathbf{y}). \tag{69}$$

By (48), the left-hand side is directly proportional to the heat flow. Its irreversible part is the algorithmic entropy production or *EP cost* $\Delta_i K$. The reversible part is the drop in base system entropy $-\Delta K$; by analogy to its ensemble-based analog in the stochastic thermodynamics literature [43,85], we call it the algorithmic *Landauer cost*. Thus, heat flow is directly proportional to the sum of EP and Landauer costs.

Equation (69) helps to clarify some common misconceptions regarding Landauer's principle [43]. For example, logical irreversibility need not result in a Landauer cost, nor in heat flow [except in the deterministic case, where (61) holds]. Sagawa [41] demonstrates this with a physical example, though a simpler example suffices: consider a finite-state Markov chain with the transition matrix $P(y,\ x) := 1/|\mathcal{X}|$. Each iteration is logically irreversible, overwriting the previous value with an independent uniformly distributed value. Nonetheless, the algorithmic entropy usually stays near $\log_2 |\mathcal{X}|$, so the Landauer cost is zero. Since P is doubly stochastic, it can be implemented without contacting a reservoir, so the heat flow is also zero.

Another common mistake is to identify the Landauer cost with a system's long-term energy consumption. In reality, Landauer costs are reversible: for a computer memory whose entropy is bounded from both above and below, positive and negative Landauer costs must approximately balance each other in the long run. That is, we have the long-run homeostasis condition $\Delta K \approx 0$, which by (69) implies $-\Delta_e K \approx \Delta_i K$. By (48), the long-term energy consumption is therefore proportional to the irreversible EP cost:

$$Q \approx k_B T \ln 2\ \Delta_i K. \tag{70}$$

This suggests an interesting accounting trick. The obvious way to compute the net heat flow Q over a long series of events is to sum the heat emission or absorption from each individual event; such a sum may include redundant Landauer costs that cancel due to opposite signs. Alternatively, (70) says that we can sum the EP costs of the individual events, and divide by $k_B T \ln 2$. The latter methodology not only avoids redundant cancellations, but also expresses the energy cost directly in terms of the algorithmic information-theoretic quantity $\Delta_i K$.

To gain some further intuition, we now examine three common types of information process: randomization, computation, and measurement. In each case, we ask whether there is an EP cost, and how that translates to net heat emission.

First, consider the generation of random data, perhaps to serve as a seed for a randomized computation. K increases, corresponding to a negative Landauer cost. In principle, Bennett [48] shows that entropy can be reversibly extracted from a heat reservoir, cooling it while randomizing a piece of digital memory. Later, the memory's entropy can be reversibly returned to the reservoir for a positive Landauer cost, warming it while clearing the memory. This cycle has zero net heat flow. Notice that the cooling step is essential: if the memory collects entropy in an uncontrolled manner, without cooling the reservoir, then (69) requires the negative Landauer cost to be offset by a positive EP cost. By the time the memory is cleared, there will be some net heat emission, as predicted by (70).

In the second case, consider a long string x resulting from a deterministic computation. Although x may appear complex, in reality its entropy $K(x)$ is about as small as the program that computed it. Only when we ignore the origins of x and toss it into a stochastic reservoir, is entropy produced. Since the string was not truly random, we have $\Delta K \approx 0$. Thus, the warming of the reservoir cannot be attributed to Landauer cost and is in fact an EP cost. In principle, a *reversible computer* can avoid EP and heat emission, clearing x by running its computation in reverse [44,86–88].

Finally, consider sensing or measurement. In Sec. V B we saw that a memory can reversibly take a measurement of another object. By interacting again with the same object, the measurement can be undone at zero cost. On the other hand, if either copy of the data is lost *without* them interacting, either because the memory is overwritten or the source object changes state, then there is an EP cost.

To see this, consider two arbitrary systems (e.g., a memory and some object), in the respective states x and y. Using (14), their total entropy decomposes as

$$K(x,\, y) = K(x) + K(y) - I(x : y).$$

As a result, the entropy production is a sum of three terms:

$$\Delta_i K := \Delta K(x,\, y) = \Delta K(x) + \Delta K(y) + \Delta(-I(x : y)). \tag{71}$$

Substituting into (44) (which again follows from Theorem 7), we obtain a formulation of the second law that explicitly accounts for the change in mutual information between the systems. It is an algorithmic analog of the result by Sagawa and Ueda [89].

Now, suppose the two systems do not interact, making each of them an isolated system. Then Theorem B 2 applies to each system individually: up to minor fluctuations, it says that their respective entropies $K(x)$ and $K(y)$ are nondecreasing. Theorem B 3 also applies, saying that $I(x : y)$ is nonincreasing. Since all three terms in the decomposition (71) are nonnegative, we can view each of them as separate EP costs. In particular, we conclude that for noninteracting systems, discarded mutual information is a form of EP.

Whether our aim is to randomize, to compute, or to measure, the absence of entropy is a resource to be carefully managed. In the first case, it is exchanged with a heat reservoir; in the second, it is encrypted by a computation; and in the third, it is stored in the mutual information between two systems. In principle, all these manipulations can be done reversibly, at zero net cost. However, when there is a mismatch between our technological mechanism and the information that it processes, then we pay an EP cost, which ultimately turns to heat according to (70).

D. An information engine

In the physical world, energy is conserved. When a system "consumes" energy, the total energy does not decrease; instead, it transforms into waste heat. Here (70) equates the heat Q with the entropy production $\Delta_i K$; thus, the "resource" that is consumed is in fact the negentropy (20).

While the utility of negentropy is apparent throughout the engineering disciplines, it is helpful to see why negentropy is useful from a purely information-theoretic point of view. To do so, we model an information-theoretic analog of a heat engine. Our "information engine" operates in an abstract Universe of coarse-grained subsystems, with no concept of reservoirs, energy, or heat. Setting $\pi := \sharp$ and identifying the state space of each subsystem with the set of binary strings of a fixed length, the negentropy (20) of any given state x reduces to approximately $|x| - K(x)$, the compressibility of x. Thus, compressible strings are the resource which should power the engine.

Let the engine have an internal memory system with state space $\mathcal{B}^m$. Using the self-delimiting encodings (12) and $\bar{x} := \overline{|x|}x$, any string $x \in \mathcal{B}^*$ that satisfies $|\bar{x}| \leqslant m$ can be encoded in memory as the concatenation of $\overline{|x|}$, x, and a padding of $m - |\bar{x}|$ zeros, which we denote by

$$e(x) := \bar{x}0^{m-|\bar{x}|} = \overline{|x|}x0^{m-|\bar{x}|} \in \mathcal{B}^m.$$

The self-delimiting prefix $\overline{|x|}$ makes $e(x)$ uniquely decodable into its three parts.

The engine uses a fixed *lossless compression algorithm:* a computable injective function $f : \mathcal{B}^* \to \mathcal{B}^*$, whose worst-case blowup

$$c := \max_{x \in \mathcal{B}^*} \{|\overline{f(x)}| - |\bar{x}|\}$$

is much less than m. Since e and f are injective, the mapping

$$e(x) \mapsto e(f(x)), \tag{72}$$

defined on the range of e, is also injective.

If f is not simply describable, we can take it to be programed onto a read-only section of memory, acting as a control in the simply describable joint mapping $g : (f,\, x) \mapsto$

TABLE I. One cycle of an information engine's operation. Its capacity is 120 bits, written as 20 base 64 characters. Incompressible strings are represented by randomly generated characters. The length prefix and the original copy of the genome are not shown. Every action is reversible.

Action	Information engine	Environment segment
Begin	`PG18Q000000000000000`	`1ksajddSG45VYummyAlphabetSoup`
Burn/clear	`PG18Q1ksajddSG45V000`	`000000000000YummyAlphabetSoup`
Reproduce	`PG18Q1ksajddSG45V000`	`CopyOfGenomeYummyAlphabetSoup`
Eat	`YummyAlphabetSoup000`	`CopyOfGenomePG18Q1ksajddSG45V`
Digest	`WAiKV000000000000000`	`CopyOfGenomePG18Q1ksajddSG45V`

$(f,\ f(x))$. Thus, no generality is lost in assuming $K(f) \stackrel{+}{=} 0$, which implies

$$K(x) \stackrel{+}{=} K(f(x)) \stackrel{+}{<} |\overline{f(x)}|.$$

Therefore, $K(x)$ sets an optimistic bound on how well the mapping (72) compresses the data in the memory. When the compression succeeds, the zero padding lengthens.

We are now ready to describe the engine's operation. Compressible strings are its fuel, to be collected from the environment, while incompressible strings are waste, to be expelled into the environment. The engine cycles between three modes:

(1) Consume ("burn") zeros to perform some task, producing waste.

(2) Expel waste, and gather ("eat") fresh fuel in its place.

(3) Refine ("digest") fuel, producing zeros and a waste byproduct.

The corresponding transitions to the memory state are summarized in a diagram:

$$e(x) \xrightarrow{\text{burn}} e(xy) \xrightarrow{\text{eat}} e(z) \xrightarrow{\text{digest}} e(f(z)).$$

Here x is a small string, perhaps compressed from the previous cycle. Hence, $e(x)$ has a large zero padding; we will see shortly how zeros are used to perform useful tasks. These tasks replace a portion of the padding with some other string y. If the engine expects x and y to be incompressible, it treats them as waste. The second stage identifies a promising location in the environment, where compressible strings might be found. With one reversible swap, the engine expels xy and gathers the (hopefully) compressible string z in its place, with $|z| = |x| + |y|$. Finally, the third stage refines the fuel z by compressing it, yielding additional zeros alongside the byproduct $f(z)$, which takes the role of x when the cycle resets.

The zeros have many uses. One is that they pay for the processing of bad fuel: if the string z turns out not to be compressible after all, then $f(z)$ may actually be longer than z, overwriting up to c of the zeros. If this happens so often as to fully deplete the supply of zeros, the engine's behavior becomes ill-defined; in that event, we consider it to have "starved to death."

Otherwise, zero padding serves as a source of ancilla bits, fueling irreversible (many-to-one) operations by embedding them as reversible (one-to-one) operations [83]. Irreversible operations include data overwrites, error correction, healing, and repair: each of these maps a larger number of "bad" states to a smaller number of "good" states. We saw an example of this in Sec. V B, where the memory of Maxwell's demon is the ancilla that enables a transition $(0,\ x) \mapsto (m(x),\ y)$, even when the second law forbids directly mapping $x \mapsto y$. Bennett [48] offers another example based on adiabatic demagnetization, consuming zeros to turn heat into work.

A living organism can use an information engine to support its growth and reproduction. A direct implementation of these operations would be irreversible, because they overwrite parts of the environment with copies of the organism's data [90]. To get a reversible implementation, the engine can absorb the data that would be overwritten, into its zero padding. Table I illustrates such an organism's operation of an information engine, for one burn-eat-digest cycle. First, it burns some of the zero padding to perform a useful function: in this case, swapping zeros onto a desired target location in the environment. Now that the target location is cleared, the organism can reversibly copy any data, such as a genome, onto it. At this point, the zero padding is almost used up. In order for the engine to recharge, it swaps in the compressible string `YummyAlphabetSoup` from the environment. Compressing this string restores the padding to a more useful length, maintaining a kind of internal homeostasis.

We leave comparisons with real-world heat and information engines, such as those studied by Leighton *et al.* [91], to future work. A fuller analogy might assign energy values to memory states, similar to the combinatorial reservoirs of Baumeler *et al.* [92] and Ebtekar [56].

VI. DISCUSSION

In order to develop ensemble-free definitions of thermodynamic quantities, we assembled ideas from stochastic thermodynamics, dynamical systems, and algorithmic information theory. The assumption of a Markovian coarse graining reduces physical systems to time-homogeneous discrete-state Markov processes. In this setting, stochastic thermodynamics defines the stochastic and Gibbs-Shannon entropies in terms of probabilistic ensembles of physical states [7]. In many instances, these are good practical approximations of the algorithmic entropy.

To deal with cases where a suitable ensemble description is not available, we propose that thermodynamics be based on the algorithmic entropy of *individual* states. Levin's [24] randomness conservation law then leads to a nonequilibrium generalization of the second law of thermodynamics (Corollary 2). To ensure the accuracy of our conclusions, we carefully accounted for some ways that information can "leak," such as the elapsed time (which allows for Poincaré

recurrence), and algorithmically complex dynamics (implemented by an exogenous control).

In terms of applications, we found that the algorithmic second law streamlines the analysis of Maxwell's demon, paving the way for thermodynamic analyses of all systems lacking a natural ensemble description. We followed up with an AIT perspective on Landauer and EP costs, which we hope will encourage more research in energy-efficient computing. In particular, (70) equates the long-term net heat flow with algorithmic entropy production.

Information-theoretic perspectives on thermodynamics are gaining traction in the physics of computing [85], biology [91], and microscopic devices more generally [45]. In light of the growing focus on fluctuation theorems [14], we hope to find more applications for the algorithmic fluctuation inequalities derived from Theorem B 1. These include the Zurek-Kolchinsky inequality (45), the Jarzynski inequality (58), and the Landauer inequality (59). In addition, future work might derive algorithmic versions of stochastic thermodynamics results not studied in this article, such as uncertainty relations and speed limit theorems [93].

While this article focuses on the second law of thermodynamics, Markov processes are known to satisfy additional information-theoretic laws. We briefly made use of the information nonincrease law (Theorem B 3), which likewise has a probabilistic version ([18], Sec. 2.8). An interesting consequence of this law is that any mutual information between systems in the present is traceable to a common cause in the past. This is a time-reversal asymmetry, perhaps even as fundamental as the second law; it may help us understand the perceptual, psychological, epistemic, and causal aspects of the so-called *arrow of time* [4,56,94,95].

Causality here is meant not in the time-symmetric sense commonly associated with Einstein's relativity, but in the asymmetric sense of Reichenbach [96], Lewis [97,98], and Bell [99,100], later refined by Pearl [101]. It generalizes the Markov property to nonlinear causal topologies. In the language of physics, the causal Markov property constrains which spacetime regions X and Y may be statistically correlated, conditional on a third region Z. Janzing and Schölkopf [102] present an algorithmic causal Markov property, while Lorenz [103] and the references therein propose quantum versions. Causal modeling describes interactions between open systems. Ito and Sagawa [104] apply it to information thermodynamics; future work might extend this using AIT.

There is yet another general law to consider. In an effort to capture the complexity of intricate structures found in living organisms, Bennett [94,105] defines the *logical depth* of x, at significance level s, to be the minimum runtime among programs, of length up to $K(x) + s$, that output x. He proves that the logical depth, if it increases, can only do so slowly. Thus, logically deep objects, such as genomes, are only created by gradual processes over a long span of time. Unlike entropy, which is maximized in the late Universe, we expect that logical depth is maximized at *intermediate* times: late enough for its gradual accumulation, but not so late as to be destroyed by heat death [106–108]. Note that both mutual information and logical depth describe ways in which the negentropy of a system becomes difficult to extract, demanding that separated systems be reunited in the former case, and that a long computation be rewound in the latter.

In future work, it would be interesting to study the interactions between entropy nondecrease, mutual information nonincrease, logical depth slow increase, and any related laws that are as yet undiscovered. Together, they seem to characterize the arrow of time, mediating the role of information in physics, computation, and intelligent life [94]. In light of the known connections between data compression, inductive learning, and intelligence [60,61,71], it might be interesting to study intelligent agent behavior from the perspective of optimizing information engines along the lines of Sec. V D.

Finally, extending algorithmic thermodynamics to incorporate quantum information remains a wide open problem. As a promising start, several quantum analogues of the description complexity have been proposed, each with different properties [109–112]. Just as chaos makes classical systems probabilistic (see Appendix A), decoherence makes quantum systems behave like mixed channels ([16], Sec. 6.2 and [38,40,113,114]), which might help explain their irreversibility.

ACKNOWLEDGMENTS

We are thankful for Xianda Sun's review of an early draft. In addition, A.E. benefited from discussions with Jason Li, Gordon Walter Semenoff, Alexander Shen, Laurent Bienvenu, Sven Bachmann, William Unruh, Danica Sutherland, William Evans, David Wakeham, and Daniel Abolafia, as well as Matthew Leighton and other participants at David Wolpert and Gülce Kardeş's workshop on Stochastic Thermodynamics and Computer Science Theory at the Santa Fe Institute. This research received no external funding.

APPENDIX A: A MARKOVIAN COARSE GRAINING

No discussion of the second law would be complete without addressing the fundamental modeling assumptions responsible for the asymmetry between past and future. To simplify matters, consider the doubly stochastic case, corresponding to a phase space partitioned into cells of equal Liouville measure. The time reversal of a *deterministic* doubly stochastic process is again doubly stochastic; by Corollary 1, its entropy can neither increase nor decrease at an appreciable rate.

Randomness breaks the symmetry: given a time-homogeneous doubly stochastic process, its time reversal need not be time-homogeneous nor double stochastic [115]. This matches our real-life macroscopic experience, where forward evolutions follow localized statistical laws, but backward evolutions do not. For example, a glass vase in free fall will shatter at a predictable time; and while the final arrangement of its pieces is chaotic and hard to predict, we can expect it to follow a well-defined statistical distribution. Moreover, our statistical prediction would not depend on any prior or concurrent happenings, e.g., at the neighbor's house.

In contrast, consider the time-reversed view, where we see a broken vase and want to retrodict its time of impact. It is hard to make even a meaningful statistical prediction. Our best attempt would be based on principles beyond the localized physics: for example, we might take into account a

conversation at the neighbor's house, telling of the accident. In the reverse dynamics, distant shards begin to converge simultaneously, in apparent violation of locality.

Experience suggests that time-homogeneous Markov processes, despite their asymmetry, are good models of real macroscopic systems. Meanwhile, the fundamental microscopic laws of nature are widely believed to be deterministic and CPT symmetric [1]. How, then, can nature's coarse-grained evolution violate this symmetry?

To demonstrate the plausibility of such an emergent asymmetry, we construct a system for which it occurs. Gaspard [30] defines the *multibaker map*: a deterministic time-reversible dynamical system that, when suitably coarse-grained, emulates a random walk. Altaner and Vollmer [8] generalize the multibaker map to emulate arbitrary Markov chains. To convey their idea in an easier fashion, we now present multibaker maps at an intermediate level of generality.

Fix an integer $m > 1$. We augment the coarse-grained state space $\mathcal{X}$ with a bi-infinite sequence of $\mathbb{Z}_m$-valued microvariables, so that the total fine-grained state space is $\mathcal{X} \times (\mathbb{Z}_m)^{\mathbb{Z}}$. Every individual fine-grained state can be written in the form

$$(x,\ (\ldots,\ r_{-2},\ r_{-1},\ r_0,\ r_1,\ r_2,\ \ldots)),$$

where $x \in \mathcal{X}$ is the coarse-grained part, and the $r_i \in \mathbb{Z}_m$ collect the remaining fine-grained information. Alternatively, we can rearrange the variables and punctuation into

$$(x.r_{-1}r_{-2}r_{-3}\ldots,\ 0.r_0r_1r_2\ldots).$$

The "0." here is purely symbolic. If we were to identify $\mathcal{X}$ with $\mathbb{Z}$, the latter notation is suggestive of the base m representation of a point in the two-dimensional "phase space" $\mathbb{R} \times [0, 1]$. There is an extensive literature that studies symbolic representations as proxies for continuous chaotic dynamical systems; for theory and examples, see Lind and Marcus [116].

At each discrete time step, the system evolves by a deterministic and reversible two-stage transformation. The first stage shifts all of the r_i by one position; we think of it as emulating microscopic chaos. The second stage applies a fixed bijection of $\mathcal{X} \times \mathbb{Z}_m$ to the pair (x, r_0); we think of it as emulating the coarse-grained physics. In summary:

$$\begin{aligned}
&(x.r_{-1}r_{-2}r_{-3}\ldots,\ 0.r_0r_1r_2\ldots)\\
\xrightarrow{\text{shift}}\ &(x.r_0r_{-1}r_{-2}\ldots,\ 0.r_1r_2r_3\ldots)\\
\xrightarrow{\text{transform}}\ &(x'.r_0'r_{-1}r_{-2}\ldots,\ 0.r_1r_2r_3\ldots).
\end{aligned}$$

The system's only source of randomness is its initial condition. At the start time $t = 0$, we allow x to have any chosen distribution, but require the r_i to be uniformly distributed, with all of the variables being independent. We can think of $r_i \in \mathbb{Z}_m$ as an m-sided die used to emulate a stochastic transition of x at the time $t = i$. In the coarse-grained view, where we marginalize all of the r_i, it is easy to verify that x's trajectory is a time-homogeneous doubly stochastic Markov chain, whose transition matrix entries are all multiples of $1/m$. It follows from Corollary 2 that $S_\sharp(x) = K(x)$ is nondecreasing up to minor fluctuations.

In fact, our multibaker map can emulate *all* such Markov chains, by a suitable choice of the bijection $T : (x,\ r_0) \mapsto (x',\ r_0')$. Indeed, recall that a Markov chain's distribution is uniquely determined by its initial condition and transition matrix. Since we already allow the initial distribution of x to be arbitrary, it remains only to emulate the doubly stochastic matrix P. Since its entries are multiples of $1/m$, we need only to assign each pair $x, y \in \mathcal{X}$ to each other with multiplicity $m \cdot P(y,\ x)$. One way to accomplish this is to fix any total order $<$ on $\mathcal{X}$, and let

$$T\left(x,\ i + m\sum_{z<y} P(z,\ x)\right) := \left(y,\ i + m\sum_{z<x} P(y,\ z)\right)$$
$$\forall x, y \in \mathcal{X},\ i \in \mathbb{Z}_{m\cdot P(y,\,x)}.$$

Thus, quite a diverse class of Markov chains arise as the coarse-grained part x of some multibaker map. In particular, this construction realizes every example of a doubly stochastic Markov chain in this article as a deterministic time-reversible map, by appending the microvariables r_i to its state.

The full construction by Altaner and Vollmer [8] relaxes the requirement that P be doubly stochastic, or that its entries have a common denominator m. The unpublished manuscript by Ebtekar [56] does the same in a different manner, and relaxes the requirement that the r_i be uniform and independent: provided that the fine-grained state starts with a continuous distribution, it is shown that the dynamics eventually stabilize to become Markovian, time-homogeneous, and doubly stochastic. Thus, any sufficiently smooth initial distribution may serve as Albert's [3] *Past Hypothesis*. Ebtekar and Hutter [4,56] provide further extensions to model causal interactions.

While these conclusions are only proven for variants of the multibaker maps, they are highly suggestive of techniques that we might try extending to realistic systems. Gaspard ([17], Sec. 4.8) suggests that we should seek a short-term ergodic property of the state's microscopic part, occuring on a much faster time scale than macroscopic ergodicity. Given a suitable coarse graining, the goal would be to prove fast convergence to Markovian behavior, long before the slower but better-understood convergence to maximum entropy. In this manner, we hope to establish the second law of thermodynamics as a mathematically rigorous property of real, CPT-symmetric systems.

APPENDIX B: CONSERVATION OF RANDOMNESS

To state the needed mathematical results in full generality, we allow the reference measure π to be nonstationary. Denoting the probability of each state transition $x \to y$ by $P(y,\ x)$, the algorithmic entropy production is defined by

$$S_{P\pi}(y \mid \widetilde{P}) - S_\pi(x \mid \widetilde{P}),$$

where $P\pi$, $\widetilde{P}$ and S_π are given by (4), (5), and (18), respectively.

Before going into formal proofs, we sketch some intuition for why we expect the entropy production to be positive. For notational convenience, let $x^* := (x,\ K(x))$ and $x^*_z := (x,\ K(x \mid z))$. Using (20), define the *conditional randomness deficiency*

$$d_P(y \mid x) := J_{P(\cdot,\,x)}(y \mid x^*) = \log_2 \frac{1}{P(y,\ x)} - K(y \mid x^*).$$

Omitting the conditioning on $\widetilde{P}$ for brevity, the algorithmic entropy production during a state transition $x \to y$ can be expressed as

$$\begin{aligned} S_{P\pi}(y) - S_{\pi}(x) &= K(y) - K(x) + \log_2 \frac{P\pi(y)}{\pi(x)} \\ &= K(y) - K(x) + \log_2 \frac{P(y,\, x)}{\widetilde{P}(x,\, y)} \\ &\overset{+}{=} K(y \mid x^*) - K(x \mid y^*) + \log_2 \frac{P(y,\, x)}{\widetilde{P}(x,\, y)} \\ &= d_{\widetilde{P}}(x \mid y) - d_P(y \mid x). \end{aligned}$$

From the last line's symmetry, one might guess that the entropy production is equally likely to be positive or negative. However, note that in general,

$$\Pr(y \mid x) = P(y,\, x), \quad \text{whereas } \Pr(x \mid y) \neq \widetilde{P}(x,\, y).$$

While randomness deficiencies typically satisfy $d_P(y \mid x) \approx 0$, under the mismatched backward probabilities we may have $d_{\widetilde{P}}(x \mid y) \gg 0$. Thus, the algorithmic entropy production measures the extent to which x is an "atypical" predecessor of y, when viewed as a sample from $\widetilde{P}(\cdot,\, y)$. For example, a low-entropy initial state x would be highly atypical for the doubly stochastic matrix $P(y,\, x) = \widetilde{P}(x,\, y) = 1/|\mathcal{X}|$, so the expected entropy production is high in this case.

Our formal proofs are based on the following lemma. Like the conditional randomness deficiency, $f(y,\, x)$ here can be interpreted as a conditional test of P randomness for y, given some data $g(x)$. When $f(y,\, x)$ is large, a statistician would reject the claim that y was sampled from the distribution $P(\cdot,\, x)$ ([49], Sec. 4.3.5 and [68,69]). Note that neither f nor g are required to be computable.

In the physical interpretation, we will see that $-f(y,\, x)$ generalizes the role of entropy production during a state transition $x \to y$. Lemma B 1 says that the *detailed fluctuation inequality* (B1) implies the *integral fluctuation inequalities* (B2), the *mean bounds* (B3), and the *tail bound* (B4). Thus, for any f, proving (B1) is sufficient to conclude the rest.

Lemma B 1. Let X, Y be $\mathcal{X}, \mathcal{Y}$-valued random variables, and write $P(y,\, x) := \Pr(Y = y \mid X = x)$. Let $f : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathcal{B}^*$ be any functions satisfying

$$\forall x \in \mathcal{X},\, y \in \mathcal{Y}, \quad f(y,\, x) \leqslant \log_2 \frac{1}{P(y,\, x)} - K(y \mid g(x)). \tag{B1}$$

Then

$$\langle 2^{f(Y,X)} \mid X \rangle < 1, \quad \langle 2^{f(Y,X)} \rangle < 1, \tag{B2}$$

$$\langle f(Y,\, X) \mid X \rangle < 0, \quad \langle f(Y,\, X) \rangle < 0, \tag{B3}$$

and for $\delta > 0$, with probability greater than $1 - \delta$,

$$f(Y,\, X) < \log_2 \frac{1}{\delta}. \tag{B4}$$

If instead (B1) holds with $\overset{+}{<}$, then so do (B3) and (B4), and (B2) then holds with $\overset{\times}{<}$.

Proof. Rearranging (B1),

$$\log_2 P(y,\, x) + f(y,\, x) \leqslant -K(y \mid g(x)).$$

To get (B2), apply Kraft's inequality (13):

$$\langle 2^{f(Y,X)} \mid X \rangle := \sum_{y \in \mathcal{Y}} P(y,\, X)\, 2^{f(y,X)} \leqslant \sum_{y \in \mathcal{Y}} 2^{-K(y \mid g(X))} < 1.$$

To get (B3), apply Jensen's inequality:

$$\langle f(Y,\, X) \mid X \rangle \leqslant \log_2 \langle 2^{f(Y,X)} \mid X \rangle < 0.$$

The unconditional versions of (B2) and (B3) follow from the law of total expectation. Finally, let $\delta > 0$. Markov's inequality on (B2) implies that, with probability greater than $1 - \delta$,

$$2^{f(Y,X)} < \frac{1}{\delta}.$$

Taking logarithms now yields (B4). The case with $\overset{+}{<}$ follows similarly. ■

As a first application of Lemma B 1, consider the case where X is constant, Y is distributed in proportion to some system's stationary distribution π, and $g(\cdot) := \widetilde{P}$ (i.e., g is a constant function that outputs a program computing $\widetilde{P}$). Then the right-hand side of (B1) reduces to the negentropy (20), and the conclusion (B4) agrees with (21).

Next, we turn to general nonequilibrium dynamics. We present detailed fluctuation inequalities for the change in each of the quantities K, π, S_π, and I.

Theorem B 1 (Detailed fluctuations). Let $\pi : \mathcal{X} \to \mathbb{R}^+ \setminus \{0\}$ be a measure. When they appear as side information, suppose $P : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^+$ and $\widetilde{P}$ from (5) are computable. Then, for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{B}^*$,

$$K(y \mid P) - K(x \mid P) \overset{+}{<} \log_2 \frac{1}{P(y,\, x)} - K(x \mid y^*_P, P), \tag{B5}$$

$$\log_2 \pi(x) - \log_2 P\pi(y) \overset{+}{<} \log_2 \frac{1}{P(y,\, x)} - K(x \mid y, \widetilde{P}), \tag{B6}$$

$$S_\pi(x \mid \widetilde{P}) - S_{P\pi}(y \mid \widetilde{P}) \overset{+}{<} \log_2 \frac{1}{P(y,\, x)} - K(y \mid x^*_{\widetilde{P}}, \widetilde{P}), \tag{B7}$$

$$I(y : z \mid P) - I(x : z \mid P) \overset{+}{<} \log_2 \frac{1}{P(y,\, x)} - K(y \mid (x, z)^*_P, P). \tag{B8}$$

Proof. For each x, $P(\cdot,\, x)$ is a probability measure computable by a constant-sized program along with (x, P); similarly, $\widetilde{P}(\cdot,\, y)$ can be computed using $(y, \widetilde{P})$. Hence, (16) implies

$$K(y \mid x, P) \overset{+}{<} \log_2 \frac{1}{P(y,\, x)}, \quad K(x \mid y, \widetilde{P}) \overset{+}{<} \log_2 \frac{1}{\widetilde{P}(x,\, y)}.$$

Now, we verify the inequalities one at a time. (B5) follows from

$$\begin{aligned} K(x \mid y^*_P, P) &\overset{+}{=} K(x, y \mid P) - K(y \mid P) \\ &\overset{+}{<} K(y \mid x, P) + K(x \mid P) - K(y \mid P) \\ &\overset{+}{<} \log_2 \frac{1}{P(y,\, x)} + K(x \mid P) - K(y \mid P). \end{aligned}$$

Similarly, (B6) follows from

$$K(x \mid y, \widetilde{P}) \overset{+}{<} \log_2 \frac{1}{\widetilde{P}(x,\ y)} = \log_2 \frac{1}{P(y,\ x)} + \log_2 P\pi(y) - \log_2 \pi(x).$$

The proof of (B7) combines the steps of the previous derivations:

$$\begin{aligned} K(y \mid x^*_{\widetilde{P}}, \widetilde{P}) &\overset{+}{=} K(x, y \mid \widetilde{P}) - K(x \mid \widetilde{P}) \\ &\overset{+}{<} K(x \mid y, \widetilde{P}) + K(y \mid \widetilde{P}) - K(x \mid \widetilde{P}) \\ &\overset{+}{<} \log_2 \frac{1}{\widetilde{P}(x,\ y)} + K(y \mid \widetilde{P}) - K(x \mid \widetilde{P}) \\ &= \log_2 \frac{1}{P(y,\ x)} + S_{P\pi}(y \mid \widetilde{P}) - S_\pi(x \mid \widetilde{P}). \end{aligned}$$

Expression (B8) was first shown by Gács *et al.* [70], but we present a simpler proof based on Gács [31]. Applying the algorithmic data processing identity (15) twice,

$$\begin{aligned} K(y \mid (x, z)^*_P, P) &\overset{+}{=} K(y \mid x^*_P, P) - I(y : z \mid x^*_P, P) \\ &\overset{+}{=} K(y \mid x^*_P, P) + I(x : z|P) - I((x, y) : z|P) \\ &\overset{+}{<} \log_2 \frac{1}{P(y,\ x)} + I(x : z \mid P) - I(y : z \mid P). \end{aligned}$$

■

Theorem B 1 says that if a transition $x \to y$ occurs with substantial probability $P(y,\ x)$, then it cannot substantially *decrease* the Liouville measure π or the algorithmic entropy S_π, nor can it substantially *increase* the description complexity K or the algorithmic mutual information I (with respect to any fixed object z). This does *not* imply that the quantities trend monotonically, since a large number of low-probability transitions may still sum to a high probability.

For example, consider a Markov chain that alternates between a "hub" state, and a uniformly random selection among a large number m of other states. Formally, let $\mathcal{X} := \mathbb{Z}_{m+1}$; for $x, y = 1, \ldots, m$, let

$$\pi(0) := m, \quad \pi(x) := 1, \quad P(0,\ 0) := 0,$$
$$P(y,\ 0) := 1/m, \quad P(0,\ x) := 1,\ P(y,\ x) := 0.$$

Then P is π-stochastic. The hub state has $K(0 \mid \widetilde{P}) \overset{+}{=} 0$, while most of the other states have $K(x \mid \widetilde{P}) \overset{+}{=} \log_2 m$. Therefore, both π and K are highly nonmonotonic, taking turns alternating between a much lower and a much higher value.

On the other hand, it is easy to check that the hub state, as well as most of the other states, have approximately $\log_2 m$ entropy. There are a few states with low entropy: for example, $S_\pi(1 \mid \widetilde{P}) \overset{+}{=} \log_2 \pi(1) = 0$. If we start from such a state, the entropy will immediately increase to about $\log_2 m$, and we will seldom return to these rare low-entropy states.

In general, S_π and I may fluctuate a bit but, unlike π and K, they trend monotonically. The nondecrease law for S_π is known as *randomness conservation*, while the nonincrease law for I is called *information nonincrease*. Both were first shown by Levin [24]. Our statement and proof take after the more pedagogical exposition of Gács [31], though we make additional changes. One is that we have extracted Lemma B 1 as a general tool, to derive these integral fluctuation inequalities from their detailed counterparts. Another is that we condition on the dual matrix $\widetilde{P}$, to eliminate some error terms from the older results. To eliminate $\widetilde{P}$ altogether, see Sec. IV E.

Note that in the main body of this article, we always have $\mathcal{X} = \mathcal{Y}$ and $P\pi = \pi$. Using (5), it follows that $\widetilde{P}$ is computable if both π and P are.

Theorem B 2 (Randomness conservation). Let $\pi : \mathcal{X} \to \mathbb{R}^+ \setminus \{0\}$ be a measure and X, Y be $\mathcal{X}, \mathcal{Y}$-valued random variables. Suppose $\widetilde{P}$, defined in terms of $P(y,\ x) := \Pr(Y = y \mid X = x)$ by (5), is computable. Then

$$\left\langle 2^{S_\pi(X|\widetilde{P}) - S_{P\pi}(Y|\widetilde{P})} \right\rangle \overset{\times}{<} 1.$$

Therefore, for $\delta > 0$, with probability greater than $1 - \delta$,

$$S_\pi(X \mid \widetilde{P}) - S_{P\pi}(Y \mid \widetilde{P}) \overset{+}{<} \log_2 \frac{1}{\delta}.$$

Proof. Let $f(x,\ y) := S_\pi(x \mid \widetilde{P}) - S_{P\pi}(y \mid \widetilde{P})$ and $g(x) := (x^*_{\widetilde{P}}, \widetilde{P})$. Then (B7) from Theorem B 1 implies that the hypotheses of Lemma B 1 hold, and therefore so do its conclusions. ■

Finally, physical applications motivate us to frame the information nonincrease law in terms of two independently evolving systems.

Theorem B 3 (Information nonincrease). For $i = 1, 2$, let $P_i : \mathcal{Y}_i \times \mathcal{X}_i \to \mathbb{R}^+$ be computable stochastic matrices, and X_i, Y_i be $\mathcal{X}_i, \mathcal{Y}_i$-valued random variables, such that for all $x_i \in \mathcal{X}_i$ and $y_i \in \mathcal{Y}_i$,

$$\Pr((Y_1, Y_2) = (y_1, y_2) \mid (X_1, X_2) = (x_1, x_2)) = P_1(y_1,\ x_1) P_2(y_2,\ x_2).$$

Then, writing $P := (P_1,\ P_2)$,

$$\left\langle 2^{I(Y_1:Y_2|P) - I(X_1:X_2|P)} \right\rangle \overset{\times}{<} 1.$$

Therefore, for $\delta > 0$, with probability greater than $1 - \delta$,

$$I(Y_1 : Y_2 \mid P) - I(X_1 : X_2 \mid P) \overset{+}{<} \log_2 \frac{1}{\delta}.$$

Proof. The pair P can be affixed with a constant-sized instruction to compute either P_1 or P_2. Applying (B8) from Theorem B 1 twice, first with $z = x_2$ and then with $z = y_1$:

$$\begin{aligned} &I(y_1 : x_2 \mid P) - I(x_1 : x_2 \mid P) \\ &\quad \overset{+}{<} \log_2 \frac{1}{P_1(y_1,\ x_1)} - K(y_1 \mid (x_1, x_2)^*_P, P), \\ &I(y_1 : y_2 \mid P) - I(y_1 : x_2 \mid P) \\ &\quad \overset{+}{<} \log_2 \frac{1}{P_2(y_2,\ x_2)} - K(y_2 \mid (y_1, x_2)^*_P, P) \\ &\quad \overset{+}{<} \log_2 \frac{1}{P_2(y_2,\ x_2)} - K(y_2 \mid (y_1, x_1, x_2)^*_P, P). \end{aligned}$$

Summing these inequalities yields

$$I(y_1 : y_2 \mid P) - I(x_1 : x_2 \mid P) \overset{+}{<} \log_2 \frac{1}{P_1(y_1,\ x_1)P_2(y_2,\ x_2)} - K(y_1, y_2 \mid (x_1, x_2)^*_P, P).$$

Let $f((x_1, x_2), (y_1, y_2)) := I(y_1 : y_2 \mid P) - I(x_1 : x_2 \mid P)$ and $g((x_1, x_2)) := ((x_1, x_2)^*_P, P)$. The desired result now follows from Lemma B 1. ■

[1] R. Lehnert, Symmetry **8**, 114 (2016).
[2] P. C. W. Davies, *The Physics of Time Asymmetry* (University of California Press, Oakland, CA, 1977).
[3] D. Z. Albert, *Time and Chance* (American Association of Physics Teachers, Maryland, 2001).
[4] A. Ebtekar and M. Hutter, Entropy **26**, 776 (2024).
[5] N. Shiraishi and K. Saito, J. Stat. Phys. **174**, 433 (2019).
[6] C. Maes, SciPost Phys. Lect. Notes 32 (2021).
[7] N. Shiraishi, *An Introduction to Stochastic Thermodynamics: From Basic to Advanced* (Springer Nature, Berlin, 2023), Vol. 212.
[8] B. Altaner and J. Vollmer, arXiv:1212.4728, based on http://dx.doi.org/10.53846/goediss-4985.
[9] G. Nicolis and C. Nicolis, Phys. Rev. A **38**, 427 (1988).
[10] C. Werndl, Stud. Hist. Philos. Sci. Part B **40**, 232 (2009).
[11] P. Figueroa-Romero, K. Modi, and F. A. Pollock, Quantum **3**, 136 (2019).
[12] P. Figueroa–Romero, F. A. Pollock, and K. Modi, Commun. Phys. **4**, 127 (2021).
[13] P. Strasberg, A. Winter, J. Gemmer, and J. Wang, Phys. Rev. A **108**, 012225 (2023).
[14] U. Seifert, Rep. Prog. Phys. **75**, 126001 (2012).
[15] L. Peliti and S. Pigolotti, *Stochastic Thermodynamics: An Introduction* (Princeton University Press, Princeton, NJ, 2021).
[16] T. Sagawa, *Entropy, Divergence, and Majorization in Classical and Quantum Thermodynamics* (Springer Nature, Berlin, 2022), Vol. 16.
[17] P. Gaspard, *The Statistical Mechanics of Irreversible Phenomena* (Cambridge University Press, Cambridge, UK, 2022).
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, Hoboken, NJ, 2006).
[19] D. Deutsch, *A Computable Universe: Understanding and Exploring Nature as Computation* (World Scientific, Singapore, 2013), pp. 551–565.
[20] D. Janzing and P. Wocjan, Open Syst. Inf. Dyn. **25**, 1850016 (2018).
[21] A. Kolchinsky and D. H. Wolpert, Phys. Rev. Res. **2**, 033312 (2020).
[22] D. H. Wolpert, arXiv:2404.16050.
[23] P. Gács, *Proceedings Workshop on Physics and Computation. PhysComp '94, Dallas* (IEEE, New York, 1994), pp. 209–216.
[24] L. A. Levin, Inf. Cont. **61**, 15 (1984).
[25] A. Kolchinsky, Phys. Rev. E **108**, 034101 (2023).
[26] C. Jarzynski, Annu. Rev. Condens. Matter Phys. **2**, 329 (2011).
[27] R. Landauer, IBM J. Res. Dev. **5**, 183 (1961).
[28] D. Janzing, R. Chaves, and B. Schölkopf, New J. Phys. **18**, 093052 (2016).
[29] B. Saussol, Rev. Math. Phys. **21**, 949 (2009).
[30] P. Gaspard, J. Stat. Phys. **68**, 673 (1992).
[31] P. Gács, arXiv:2105.04704.
[32] K. Maruyama, F. Nori, and V. Vedral, Rev. Mod. Phys. **81**, 1 (2009).
[33] G. Carcassi and C. A. Aidala, Stud. Hist. Philos. Sci. Part B **71**, 60 (2020).
[34] W. H. Zurek, Phys. Rev. A **40**, 4731 (1989).
[35] P. Ehrenfest and T. Ehrenfest, *The Conceptual Foundations of the Statistical Approach in Mechanics* (Courier Corporation, North Chelmsford, MA, 1959).
[36] V. Latora and M. Baranger, Phys. Rev. Lett. **82**, 520 (1999).
[37] M. Einsiedler, E. Lindenstrauss, and T. Ward, *Entropy in Ergodic Theory and Topological Dynamics* (2021), https://tbward0.wixsite.com/books/entropy.
[38] W. H. Zurek, Phys. Scr. **T76**, 186 (1998).
[39] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, Phys. Rev. Lett. **98**, 080602 (2007).
[40] M. Esposito, K. Lindenberg, and C. V. den Broeck, New J. Phys. **12**, 013013 (2010).
[41] T. Sagawa, Prog. Theor. Phys. **127**, 1 (2012).
[42] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, Nat. Phys. **11**, 131 (2015).
[43] D. H. Wolpert, J. Phys. A: Math. Theor. **52**, 193001 (2019).
[44] Ä. Baumeler and S. Wolf, Phys. Rev. E **100**, 052115 (2019).
[45] J. M. R. Parrondo, J. Tabanera-Bravo, F. Fedele, and N. Ares, *European ESiM Workshop "Crossroad of Maxwell Demon"* (Springer, Berlin, 2023), pp. 1–31.
[46] G. Manzano, G. Kardeş, É. Roldán, and D. H. Wolpert, Phys. Rev. X **14**, 021026 (2024).
[47] R. Frigg and C. Werndl, The Monist **102**, 424 (2019).
[48] C. H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982).
[49] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 4th ed. (Springer, Cham, Texts in Computer Science, 2019).
[50] M. P. Frank, arXiv:physics/0506128.
[51] D. Newell and E. Tiesinga, *The International System of Units (SI)*, 2019 Edition, Special Publication (NIST SP) (National Institute of Standards and Technology, Gaithersburg, MD, 2019).
[52] C. A. Bédard, S. Berthelette, X. Coiteux-Roy, and S. Wolf, arXiv:2401.12119.
[53] J. Baez and M. Stay, Math. Struct. Comp. Sci. **22**, 771 (2012).
[54] A. N. Kolmogorov, Math. Ann. **112**, 155 (1936).
[55] M. Nagasawa, Nagoya Math. J. **24**, 177 (1964).
[56] A. Ebtekar, arXiv:2109.09709.
[57] L. Boyle, K. Finn, and N. Turok, Phys. Rev. Lett. **121**, 251301 (2018).
[58] L. Boyle, K. Finn, and N. Turok, Ann. Phys. **438**, 168767 (2022).
[59] É. Roldán, I. Neri, R. Chetrite, S. Gupta, S. Pigolotti, F. Jülicher, and K. Sekimoto, Adv. Phys. **72**, 1 (2023).

[60] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability* (Springer Science & Business Media, Berlin, 2004).
[61] M. Hutter, D. Quarel, and E. Catt, *An Introduction to Universal Artificial Intelligence* (CRC Press, Boca Raton, FL, 2024).
[62] P. Grünwald and P. M. B. Vitányi, arXiv:cs/0410002.
[63] S. Rathmanner and M. Hutter, Entropy **13**, 1076 (2011).
[64] L. A. Levin, in *Doklady Akademii Nauk* (Russian Academy of Sciences, Moscow, Russia, 1976), Vol. 227, pp. 33–35.
[65] N. K. Vereshchagin and P. M. Vitányi, IEEE Trans. Inf. Theory **50**, 3265 (2004).
[66] P. Gács, Theor. Comput. Sci. **341**, 91 (2005).
[67] V. Vovk, in *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday* (Springer, Berlin, 2020), pp. 323–340.
[68] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer, Stat. Sci. **38**, 576 (2023).
[69] A. Ramdas and R. Wang, arXiv:2410.23614.
[70] P. Gács, J. T. Tromp, and P. M. B. Vitányi, IEEE Trans. Inf. Theory **47**, 2443 (2001).
[71] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length* (Springer, Berlin, 2005).
[72] N. Vereshchagin and A. Shen, in *Computability and Complexity: Essays Dedicated to Rodney G. Downey on the Occasion of His 60th Birthday* (Springer, Berlin, 2016), pp. 669–737.
[73] G. J. Chaitin, J. ACM **21**, 403 (1974).
[74] R. J. Solomonoff, in *Information Theory and Statistical Learning*, edited by F. Emmert-Streib and M. Dehmer (Springer, Boston, MA. 2009), pp. 1–23.
[75] D. H. Wolpert, in *Soft Computing and Industry: Recent Applications*, edited by R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, and F. Hoffmann (Springer, London, 2002), pp. 25–42.
[76] M. P. Müller, Quantum **4**, 301 (2020).
[77] M. Austern and A. Maleki, IEEE Trans. Inf. Theory **66**, 1232 (2019).
[78] A. N. Kolmogorov, Russ. Math. Surv. **38**, 29 (1983).
[79] E. A. Guggenheim, J. Chem. Phys. **7**, 103 (1939).
[80] W. H. Zurek, Nature (London) **341**, 119 (1989).
[81] P. Révész, Acta Math. Acad. Sci. Hung. **13**, 187 (1962).
[82] M. S. Neyman, Telecommunications and Radio Engineering-USSR, 68 (1966).
[83] T. Sagawa, J. Stat. Mech. (2014) P03025.
[84] M. P. Frank, in *International Conference on Reversible Computation* (Springer, Berlin, 2018), pp. 3–33.
[85] D. H. Wolpert, J. Korbel, C. Lynn, F. Tasnim, J. Grochow, G. Kardeş, J. Aimone, V. Balasubramanian, E. de Giuli, D. Doty *et al.*, Proc. Natl. Acad. Sci. USA **121**, e2321112121 (2024).
[86] P. M. B. Vitányi, in *Proceedings of the 2nd Conference on Computing Frontiers* (ACM Press, New York, NY, 2005), pp. 435–444.
[87] E. D. Demaine, J. Lynch, G. J. Mirano, and N. Tyagi, in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science* (ACM Press, New York, NY, 2016), pp. 321–332.
[88] K. Morita, *Theory of Reversible Computing*, Monographs in Theoretical Computer Science. An EATCS Series (Springer, Tokyo, 2017).
[89] T. Sagawa and M. Ueda, Phys. Rev. Lett. **109**, 180602 (2012).
[90] S. D. Devine, Biosystems **140**, 8 (2016).
[91] M. P. Leighton, J. Ehrich, and D. A. Sivak, Phys. Rev. X **14**, 041038 (2024).
[92] Ä. Baumeler, C. Rieger, and S. Wolf, in *Proceedings of the 2022 IEEE Information Theory Workshop (ITW)* (IEEE, Piscataway, NJ, 2022), pp. 362–367.
[93] V. T. Vo, T. Van Vu, and Y. Hasegawa, Phys. Rev. E **102**, 062132 (2020).
[94] C. H. Bennett, Physical Origins of Time Asymmetry, 33 (1994).
[95] D. H. Wolpert and J. Kipper, Entropy **26**, 170 (2024).
[96] H. Reichenbach, *The Direction of Time* (University of California Press, Oakland, CA, 1956), Vol. 65.
[97] D. Lewis, J. Philos. **70**, 556 (1973).
[98] D. Lewis, Noûs **13**, 455 (1979).
[99] J. S. Bell, Epistemol. Lett. **9**, 11 (1975).
[100] J. S. Bell, Epistemol. Lett. **15**, 79 (1977).
[101] J. Pearl, *Causality*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2009).
[102] D. Janzing and B. Schölkopf, IEEE Trans. Inf. Theory **56**, 5168 (2010).
[103] R. Lorenz, Synthese **200**, 424 (2022).
[104] S. Ito and T. Sagawa, Phys. Rev. Lett. **111**, 180603 (2013).
[105] C. H. Bennett, *The Universal Turing Machine: A Half-Century Survey* (Oxford University Press, 1988), pp. 227–257.
[106] L. Antunes, L. Fortnow, D. V. Melkebeek, and V. N. Variyam, Theor. Comput. Sci. **354**, 391 (2006).
[107] S. Aaronson, S. M. Carroll, and L. Ouellette, arXiv:1405.6903.
[108] K. J. Jeffery and C. Rovelli, Trends Neurosci. **43**, 467 (2020).
[109] A. Berthiaume, W. V. Dam, and S. Laplante, J. Comput. Syst. Sci. **63**, 201 (2001).
[110] P. Gács, in *Proceedings of the 16th Annual IEEE Conference on Computational Complexity* (IEEE, New York, 2001), pp. 274–283.
[111] P. M. B. Vitányi, IEEE Trans. Inf. Theory **47**, 2464 (2001).
[112] C. E. Mora, H. J. Briegel, and B. Kraus, Int. J. Quantum Inf. **05**, 729 (2007).
[113] A. Peres, Phys. Rev. A **30**, 1610 (1984).
[114] M. A. Schlosshauer, *Decoherence: And the Quantum-to-Classical Transition* (Springer Science & Business Media, New York, 2007).
[115] T. M. Cover, Physical Origins of Time Asymmetry, 98 (1994).
[116] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2021).