Limits on inferring the past

Nathaniel Rupprecht and Dervis C. Vural*
Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA

(Received 18 January 2018; revised manuscript received 10 May 2018; published 29 June 2018)

Here we define and study the properties of retrodictive inference. We derive equations relating retrodiction entropy and thermodynamic entropy, and as a special case, show that under equilibrium conditions, the two are identical. We demonstrate relations involving the Kullback-Leibler divergence and retrodiction probability, and bound the time rate of change of retrodiction entropy. As a specific case, we invert various Langevin processes, inferring the initial condition of N particles given their final positions at some later time. We evaluate the retrodiction entropy for Langevin dynamics exactly for special cases, and find that one's ability to infer the initial state of a system can exhibit two possible qualitative behaviors depending on the potential energy landscape, either decreasing indefinitely, or asymptotically approaching a fixed value. We also study how well we can retrodict points that evolve based on the logistic map. We find singular changes in the retrodictivity near bifurcations. Counterintuitively, the transition to chaos is accompanied by maximal retrodictability.

DOI: 10.1103/PhysRevE.97.062155

I. INTRODUCTION

Many astonishing facts about the origin of the universe, evolution of life, or history of civilizations will never be directly observed, but will only be inferred in the light of their manifestations in the present. Evolved forward in time, any state of knowledge, regardless of how exact, will invariably deteriorate into an entropy maximizing probability distribution [1–4]. How rapidly does our knowledge of the past, as inferred from a measurement made in the present, deteriorate, going backwards in time?

While methods exist for inferring the origin of an observed final state [5–8], or inferring some original data after it has been corrupted [9,10] we know little about how accurately the initial state of a many-body system can be characterized given its present state, how quickly a system forgets its initial state due to thermal fluctuations, and how the limit our ability to infer the past depends on system parameters. The answers to these questions should lie in nonequilibrium statistical mechanics, where thermal motion is incorporated into mechanical laws [11–13]. In systems where thermal collisions erase the information pertaining to past states of particles, the Fokker-Planck equation constitutes the groundwork of nonequilibrium analysis [14–20].

Here we determine the theoretical limits to inferring the initial state of a system, which we refer to as "retrodiction" in contrast to prediction. We quantify the quality of retrodiction in terms of retrodiction entropy, S_R. We derive a relationship between thermodynamic entropy and retrodiction entropy, and report a lower bound on its generation rate. Then, to apply these ideas to a specific problem, we consider a collection of particles coupled to a thermal bath, and obtain the time dependence of S_R in convex, concave, and flat potentials. To establish whether chaos fundamentally influences retrodictability, we also investigate the retrodiction entropy of the logistic map as it transitions from the nonchaotic regime to the chaotic regime. Finally, we conclude our discussion with a comparison of retrodiction entropy to other inverse statistical methods and methods for comparing predictability and retrodictability.

II. DEFINITIONS AND NOTATION

Our system consists of a set states Ω, a prior distribution on the set of states, P_0, and a transition probability function $\mathcal{T}$. The state space Ω will depend on the problem at hand, it could for example be the space of all possible positions and velocities of a collection of particles (i.e., phase space). The prior distribution specifies how the system will be initialized; $P_0(\alpha)$ is the probability that the system will be prepared in the state $\alpha \in \Omega$. The transition probability $\mathcal{T}(\omega|\alpha;t)$ is the probability that the system ends in the state $\omega \in \Omega$ given that it started in the state $\alpha \in \Omega$ and evolved for a time t. We will generally suppress the time variable.

The probability $\mathcal{R}(\alpha|\omega;t) = \mathcal{R}_\omega(\alpha)$ that the initial state was α given the final state ω, is given by the Bayes theorem,

$$\mathcal{R}(\alpha|\omega;t) = \frac{\mathcal{T}(\omega|\alpha;t)P_0(\alpha)}{P_t(\omega)} = \frac{\mathcal{T}(\omega|\alpha;t)P_0(\alpha)}{\sum_{\alpha'}\mathcal{T}(\omega|\alpha';t)P_0(\alpha')}. \quad (1)$$

where P_t is the prior distribution P_0 evolved forwards in time. $\mathcal{R}$ would typically be called the likelihood or the posterior distribution. In the present context, we will refer to it as the retrodiction probability, and define the entropy associated with it as the retrodiction entropy,

$$S_R(\omega) = -\sum_\alpha \mathcal{R}_\omega(\alpha)\log\mathcal{R}_\omega(\alpha). \quad (2)$$

Intuitively, the larger $S_R(\omega)$ is, the less accurately the initial state can be inferred given a measurement of the final state, ω.

Note that S_R is a function of the final state observed after a single realization of a stochastic process. If the process were to be run again, the particles would end up elsewhere, and

*Corresponding author: dvural@nd.edu

have a different S_R associated with that final state. As such, it will be useful to define S_R averaged over all possible final measurements, $\langle S_R \rangle$.

A related quantity of interest is the Kullback-Leibler (KL) divergence $D[p\|q] = \sum_x p(x) \log p(x)/q(x)$, measures the amount of overlap between two distributions $p(x)$ and $q(x)$ [21]. Thus, another useful measure of retrodictability is the KL divergence $D[\mathcal{R}_\omega \| P_0]$ between $\mathcal{R}$ and P_0, which quantifies the amount of information gained over the prior upon a measurement. As our ability to infer the past decreases, the retrodiction probability coincides more with the prior probability, the KL divergence decreases. Ultimately, $D[\mathcal{R}_\omega \| P] = 0$ as the measurement ω provides no additional information regarding the initial state beyond what we already know; the prior, P.

A. Notation

Throughout, we denote the average over all free parameters by $\langle \cdot \rangle$. However, there are two different types of averages that are indicated by this notation: averages over the distribution on initial states, and averages over distribution on final states. When we average over quantities where the free variable ranges over initial states, we use a probability weight P_0 for each such free variable. For quantities where the free variable ranges over final states, we use a probability weight P_t for each such free variable. In the case where there are multiple states that are being averaged over, we include a subscript to indicate that there is a free variable to be averaged over. For example,

$$\langle S_R \rangle = \sum_\omega P_t(\omega) S_R(\omega)$$

$$\langle S_T \rangle = \sum_\alpha P_0(\alpha) S_T(\alpha)$$

$$\langle D[\mathcal{T}_{\alpha_1} \| \mathcal{T}_{\alpha_2}] \rangle = \sum_{\alpha_1, \alpha_2} P_0(\alpha_1) P_0(\alpha_2) D[\mathcal{T}_{\alpha_1} \| \mathcal{T}_{\alpha_2}].$$

III. GENERAL PROPERTIES OF RETRODICTION

A. Relation between retrodiction and thermodynamics

To facilitate readability onwards, we expose only the crucial steps in the main text, leaving the proofs and derivations to the Appendixes.

Our first key result is the relationship between retrodiction entropy and thermodynamic entropy

$$\langle S_R \rangle = \langle S_T \rangle - (S_t - S_0). \tag{3}$$

Here $\langle S_T \rangle$ is the average entropy associated with the transition probability $\mathcal{T}_\alpha(\omega)$, whereas S_0 and S_t are the entropies associated with the prior probability P_0, and the observation probability, P_t. Eq. (3) relates our ability to infer the past, $\langle S_R \rangle$, to our ability to predict the future, $\langle S_T \rangle$ and S_t. This identity is derived in Appendix A.

Note that Eq. (3) holds for processes both in or out of equilibrium, and provides useful insights on the general properties of S_R. For short times, $P_t \simeq P_0$, so $\lim_{t\to 0+} \langle S_R \rangle / \langle S_T \rangle = 1$. For long times, if the system converges to a stationary distribution P_∞, (as is the case in a bounded space or trapping potential), then P_t and $\mathcal{T}_\alpha(\omega)$ must approach P_∞ independent of the starting state, and (3) implies $\lim_{t\to\infty} \langle S_R \rangle = S_0$, i.e., we cannot guess the initial state any better than using whatever we already knew before making the measurement.

As another interesting special case, we consider what happens if the prior probability P_0 coincides with the stationary state probability P_∞ (assuming one exists). Then $S_t = S_0$ for all times t, and (3) implies

$$\langle S_R(t) \rangle = \langle S_T(t) \rangle. \tag{4}$$

For example, if we are inferring the past of a system in equilibrium we would be drawing the initial state of the system out of the equilibrium distribution, i.e., using $P_0(s) = e^{-\beta E(s)}/Z$ as the prior probability, measure the positions of some particles, and ask where they used to be. Eq. (4) tells us that in equilibrium, the rate of thermodynamic entropy and retrodiction entropy generation is the same. Our ability to predict the future fades at exactly the same rate as our ability to infer the original state of the system.

No such correspondence need hold for nonequilibrium processes. For a system with equilibrium entropy S_{eq}, if $S_0 > S_{\text{eq}}$ then S_t will decrease from S_0 at $t = 0$ to S_{eq} as $t \to \infty$. Thus $\langle S_R \rangle > \langle S_T \rangle$. In this case, we know that particles will gather, so we know better where they will be in the future than where they were originally. In contrast, if $S_0 < S_{\text{eq}}$, S_t will increase in time and $\langle S_R \rangle < \langle S_T \rangle$. Here, we know more about where the particles were originally than where they will be in the future. To sum up, the more certain we can be about the state of the system in the future, the less certain we are about where the system started out in the past.

B. Experimental measurement of retrodictability

It is instructive to view (3) from a practical, empirical perspective. Consider a system of particles evolving in a potential energy landscape $U(\vec{x})$ while coupled to a heat bath. Can we estimate bounds on $\langle S_R \rangle$ without knowing the microscopic dynamics of the system (e.g., the interparticle interactions) or the potential energy landscape, but only using thermodynamic measurements?

This is possible under certain conditions. We can initialize a system such that particles are in state α with probability $P_0(\alpha)$, let the particles evolve for a time t, calorimetrically obtain the change in thermodynamic entropy via $\Delta S_\alpha = \int_\alpha dQ/T$, and then average this over multiple instances to obtain $\langle S_T \rangle_s$ (the sample average of entropy). The identity $dS = dQ/T$ holds when the system moves along a reversible path. While it is not trivial to measure S_t for processes out of equilibrium, we can use the equilibrium result, $\langle S_T \rangle = \langle S_R \rangle$ [Eq. (4)] and the second law, to place an upper bound on average retrodiction entropy, for any process (in or out of equilibrium),

$$\langle S_R \rangle < \langle S_T \rangle_s + S_0.$$

Under special conditions, we can do better than an inequality. If the prior distribution is uncorrelated $P_0(x_1, \ldots, x_N) = p(x_1)p(x_2)\ldots p(x_N)$, and if interactions between particles are negligible, then

$$P_t(y_1, \ldots, y_N) = \prod_{k=1}^{N} \left(\sum_{x_k} \mathcal{T}(y_k|x_k; t) p(x_k) \right) \equiv \prod_k q(y_k).$$

Since each term in this product is independent, the entropy is extensive $S_t = NH[q]$, and $S_0 = NH[p]$. Thus, an experimentalist can measure $S_t - S_0$ by placing $M \gg 1$ particles with a number density $p(x)$, allow the particles to evolve for a time t, and again calorimetrically integrate $\Delta S = \int dQ/T$ to obtain $S_t - S_0 \simeq N\Delta S/M$. Note that since $M \gg 1$, ΔS will be deterministic. Thus from (3) the retrodictability becomes a difference of two entropy measurements,

$$\langle S_R \rangle = \langle \Delta S \rangle_s - \frac{N}{M}\Delta S. \tag{5}$$

The first term on the right is measured by initializing particles individually at α with probability $P_0(\alpha)$ and averaging all outcomes, whereas the second term, by a single shot measurement of a gas initialized with density $P_0(\alpha)$. We emphasize that this experimental protocol to obtain (5) will be valid only when interparticle interactions are negligible, and for an uncorrelated prior, but as long as these assumptions hold, $\langle S_R \rangle$ can be known by only performing thermodynamic measurements, without needing to know the underlying potential or microscopic dynamics.

C. Continuous space and divergence relations

For a continuous state space, we may consider S_R to be a differential entropy, which is not invariant under a change of variables. In contrast, $D[\mathcal{R}_\omega \| P]$ is invariant under changes of variables, and therefore may be a more desirable measure. We derive, in a similar manner to (3),

$$\langle D[\mathcal{R}_\xi \| P] \rangle = S_t - \langle S_T \rangle = S_0 - \langle S_R \rangle.$$

Markovian stochastic processes are known to have a KL divergence that are nonincreasing in time [21]. Thus we are motivated to ask how the KL divergence between two forward processes $\mathcal{T}_\alpha$, compares to the KL divergence between two retrodiction probabilities $\mathcal{R}_\omega$. First, we show (cf. Appendix B)

$$\langle D[\mathcal{R}_{\omega_1} \| \mathcal{R}_{\omega_2}] \rangle = \langle D[P_t \| \mathcal{T}_\alpha] \rangle + S_t - \langle S_T \rangle$$
$$\langle D[P_t \| \mathcal{T}_\alpha] \rangle = \langle D[\mathcal{T}_{\alpha_1} \| \mathcal{T}_{\alpha_2}] \rangle + \langle S_T \rangle - S_t.$$

Combining these gives us the relationship

$$\langle D[\mathcal{R}_{\omega_1} \| \mathcal{R}_{\omega_2}] \rangle = \langle D[\mathcal{T}_{\alpha_1} \| \mathcal{T}_{\alpha_2}] \rangle. \tag{6}$$

Thus, the average amount of overlap between different retrodiction probability distributions is exactly equal to the average amount of overlap between different forward distributions (cf. Appendix B). Taking the time derivative of both sides tells us that the average rate of increase is the same for forward and reverse probabilities, and that this quantity is nonincreasing [21]. In Appendix B, we list all the KL divergence relations between the distributions $\mathcal{T}$, $\mathcal{R}$, P_0, and P_t.

D. Lower bound to retrodiction entropy generation

We can establish a lower bound on the time rate of change of retrodiction entropy in terms of forward entropies and KL divergences. Differentiating (3) and using the convexity of log gives us an upper bound on the rate of change of S_t (cf. Appendix C),

$$\dot{S}_t \leqslant \langle \dot{S}_T \rangle + \frac{\partial}{\partial t}\langle D[\mathcal{T}_{\alpha_1} \| \mathcal{T}_{\alpha_2}] \rangle - \left\langle \frac{\partial}{\partial t} D[P_0 \| \mathcal{R}_\omega] \right\rangle. \tag{7}$$

Using the theorem on Markov processes, we know that the second term in (7) is $\leqslant 0$. The last term in (7) measures the divergence between the prior state and the retrodiction probability, which should decrease with time as the reconstructed probability approaches the prior. Rearranging (7), we get,

$$\frac{\partial}{\partial t}\langle S_R \rangle \geqslant -\frac{\partial}{\partial t}\langle D[\mathcal{T}_{\alpha_1} \| \mathcal{T}_{\alpha_2}] \rangle + \left\langle \frac{\partial}{\partial t} D[P \| \mathcal{R}_\omega] \right\rangle.$$

E. Information theoretical interpretation

From an information theoretic point of view, retrodiction entropy is the amount of information required to specify which state the system was initialized, given an observation of its final state. The KL divergence between the retrodiction probability $\mathcal{R}_\omega$, and the prior distribution P_0 is a measure of how much information has been gained by making a measurement (above and beyond the information contained in the prior). The KL divergence is asymmetric in its arguments, $D[\mathcal{R}_\omega \| P_0] \neq D[P_0 \| \mathcal{R}_\omega]$. However, there is a good reason for preferring $D[\mathcal{R}_\omega \| P_0]$ over $D[P_0 \| \mathcal{R}_\omega]$. Letting X_0, X_t be the random variables for the configuration at times 0 and t, it can be shown that $\langle D[\mathcal{R}_\omega \| P_0] \rangle = I(X_0; X_t)$ where $I(\cdot,\cdot)$ is the mutual information. In other words, the average KL divergence between retrodiction probabilities and the prior is the mutual information between the initial and final states of the system. We can use this and our other formulas to write retrodiction entropy in terms of mutual information,

$$\langle S_R \rangle = S_0 - I(X_0; X_t) = H(X_0) - I(X_0; X_t). \tag{8}$$

While it is impossible to evaluate quantities such as $D[\mathcal{R}_\omega \| P_0]$ or $S_R(\omega)$ for a specific ω without being given a specific problem (and being able to evaluate the transition probabilities for that problem), Eqs. (1)–(4) and (6)–(8) hold true quite generally, for any system in or out of equilibrium.

IV. RETRODICTION OF BROWNIAN PARTICLES IN A POTENTIAL

Following these general results, we now study a specific physical system, the retrodiction entropy of Brownian particles diffusing in a potential. The αth coordinate ($\alpha = x, y, z, \ldots$) of the kth particle, will be written as $x^{(k)} = \{x_\alpha^{(k)}\}$, and for the initial state, the αth coordinate of the initial position will be written as $y = \{y_\alpha\}$. In other words, latin superscripts index particles $1, \ldots, N$ while greek subscripts indicate their coordinates, $1, \ldots, d$.

Suppose N particles are released at the same position at $t = 0$ and evolve in a potential $U(\vec{x})$ according to Langevin dynamics. The evolution of the state probability distribution $p(\vec{x},t)$ is governed by the general Fokker-Planck equation,

$$\frac{\partial p(\vec{x},t)}{\partial t} = \sum_{\alpha,\beta} \frac{\partial^2 [D_{\alpha\beta}(\vec{x},t)p(\vec{x},t)]}{\partial x_\alpha \partial x_\beta} - \sum_\alpha \frac{\partial [\mu_\alpha(\vec{x},t)p(\vec{x},t)]}{\partial x_\alpha},$$

where $\mu_\alpha(\vec{x},t)$ is a drift term and $D_{\alpha\beta}(x,t)$ is the diffusion tensor. Since particles are independent and follow identical transition rules, the probability that N particles starting at state

x, end in states $x^{(1)},\ldots,x^{(N)}$ is

$$\mathcal{T}(x^{(1)},\ldots,x^{(N)} \mid y;t) = \prod_{k=1}^{N} p(x^{(k)} \mid y;t). \tag{9}$$

The retrodiction probability $\mathcal{R}(y|x^{(1)},\ldots,x^{(N)})$ is then the probability that the initial position of the cluster of particles was y given the N observed final positions $\{x^{(k)}\}$.

A. Retrodiction entropy of a Gaussian process

Consider a process with (individual) probability distributions

$$p(x^{(k)}|y;t) = \prod_{\alpha=1}^{d} \frac{\exp\left[\left(x_\alpha^{(k)} - \lambda_\alpha(t)y_\alpha\right)^2/D_\alpha(t)\right]}{\sqrt{\pi D_\alpha(t)}}. \tag{10}$$

Here, the transition probability $\mathcal{T}(x^{(1)},\ldots,x^{(N)}|y)$ is

$$\mathcal{T} = \prod_{\alpha=1}^{d} \frac{\exp\left[-\sum_{k=1}^{N}\left(x_\alpha^{(k)} - \lambda_\alpha(t)y_\alpha\right)^2/D_\alpha(t)\right]}{[\pi D_\alpha(t)]^{N/2}} \tag{11}$$

since all particles start at x_α. Note that we allow the generalized diffusion and drift to be different in every dimension α. Suppose the prior probability for the initial position of the cluster of particles is Gaussian, centered at the origin,

$$P_0(y) = \prod_{\alpha=1}^{d} \left(2\pi\sigma_\alpha^2\right)^{-1/2} \exp\left[-y_\alpha^2/\left(2\sigma_\alpha^2\right)\right]. \tag{12}$$

The observation probability of a configuration is then

$$\begin{aligned} &P_t(x^{(1)},\ldots,x^{(N)};t) \\ &= \prod_{\alpha=1}^{d} \left(2\pi^N\sigma_\alpha^2 D_\alpha(t)^N\right)^{-1/2} \sqrt{\frac{D_\alpha(t)\kappa_\alpha(t)}{N\lambda_\alpha(t)^2}} \\ &\quad\times \exp\left[-\frac{N}{D_\alpha(t)}\left(\left\langle x_\alpha^2\right\rangle - \kappa_\alpha(t)\langle x_\alpha\rangle^2\right)\right], \end{aligned} \tag{13}$$

where, $\kappa_\alpha(t) = \{1 + D_\alpha(t)/[2N\sigma_\alpha^2\lambda_\alpha(t)^2]\}^{-1}$ and $\langle x_\alpha^n\rangle = \sum_{k=1}^{N}[x_\alpha^{(k)}]^n/N$. From this and $\mathcal{T}$, P, we can evaluate the retrodiction probability

$$\begin{aligned} &\mathcal{R}(y|x^{(1)},\ldots,x^{(N)};t) \\ &= \prod_{\alpha=1}^{d} \sqrt{\frac{N\lambda_\alpha(t)^2}{\pi D_\alpha(t)\kappa_\alpha(t)}} \\ &\quad\times \exp\left[-\left(\frac{N\lambda_\alpha(t)^2}{D_\alpha(t)\kappa_\alpha(t)}\right)\left(y_\alpha - \frac{\kappa_\alpha(t)}{\lambda_\alpha(t)}\langle x_\alpha\rangle\right)^2\right]. \end{aligned}$$

As this is a Gaussian distribution, it is straightforward to evaluate its entropy, the retrodiction entropy,

$$S_R = \frac{1}{2}\log\left[\left(\frac{\pi e}{N}\right)^d \prod_{\alpha=1}^{d} \frac{D_\alpha(t)}{\lambda_\alpha(t)^2 + D_\alpha(t)/\left(2\sigma_\alpha^2 N\right)}\right]. \tag{14}$$

Note that in the limit of $\sigma_\alpha \to \infty$ in all directions, we obtain the case of a uniform (non-normalizable) prior over all space. In this case, or in the case that σ's are finite and particles are scattered off by external forces, i.e., $\lambda_\alpha(t) \to \infty$ as $t \to \infty$, the retrodiction entropy is

$$S_R = (d/2)\log\{\pi e D_{\mathrm{GM}}(t)/[N\lambda_{\mathrm{GM}}(t)^2]\},$$

where the subscript GM indicates a geometric mean over the different directions α. The individual entropies of the distributions $\mathcal{T}$, P_0, and P_t are listed in Appendix A, which also serves to verify (3).

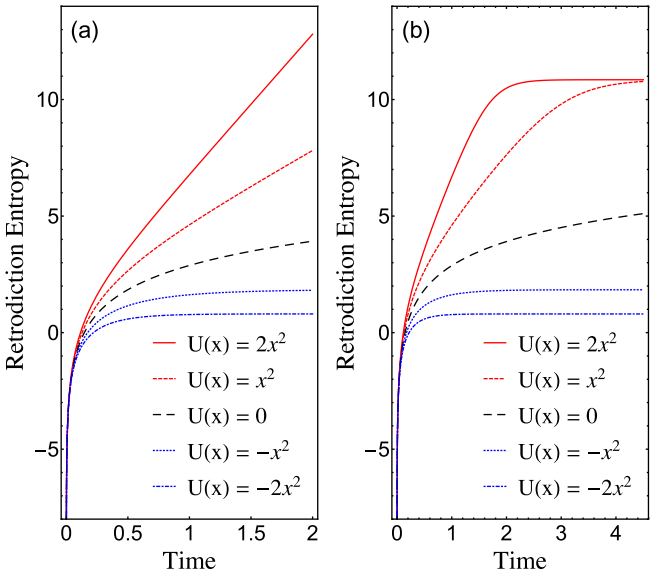

FIG. 1. Retrodiction entropy generation. The lines for potential and the labels both go from top to bottom. Left: The retrodiction entropy S_R of five particles in a convex, flat, and concave potential $U(\vec{x})$ with a uniform prior. S_R quantifies how poorly the initial state of the particles can be inferred, backwards in time. Free and trapped particles forget their origin monotonically, whereas particles dispersing in a concave potential remember their past no matter how much time passes. Right: An analogous plot, but with a Gaussian prior instead of a uniform prior. Free and trapped particles saturate to having maximum retrodiction entropy, whereas particles in a concave potential still remember their past, just as in the case of a uniform prior.

B. Convex and concave potentials

Two processes that have analytical solutions to the Fokker-Planck equation are Wiener and Ornstein-Uhlenbeck processes, describing Brownian particles in flat $U(\vec{x}) = \alpha + \vec{\beta}\cdot\vec{x}$ and parabolic $U(\vec{x}) = \alpha + \vec{\beta}\cdot\vec{x} + \theta\vec{x}^2$ potentials. We evaluate the retrodiction entropy for these special cases, and find that it diverges for particles random walking in flat and convex potentials ($\theta \geqslant 0$) indicating that the system steadily forgets its past. In contrast, concave ($\theta \leqslant 0$) potentials have a retrodiction entropy that asymptotically approach a constant less than S_P, indicating that the system always retains the memory of its initial state (see Fig. 1).

The distribution of a free Brownian particle is

$$p(x \mid y;t) = \prod_{\alpha=1}^{d}(4\pi D_\alpha t)^{-1/2}\exp[-(x_\alpha - y_\alpha)^2/4D_\alpha t].$$

In this case, the functions in (10) are $D_\alpha(t) = 4D_\alpha t$ and $\lambda_\alpha(x_\alpha) = 1$. Thus,

$$S_R = \frac{1}{2}\log\left[(4\pi e)^d \prod_{\alpha=1}^{d} \frac{\sigma_\alpha^2 D_\alpha t}{2D_\alpha t + \sigma_\alpha^2 N}\right].$$

In the limit of $\sigma_\alpha \to \infty$, S_R increases at a logarithmic rate at all times. If the σ's are finite, then at long times, $S_R \to {}^{d}/_{2}\log[4\pi e\sigma_{GM}^2]$, which is just the entropy of the prior distribution P_0. For short times, we have

$$S_R \sim (d/2)\log(4\pi e D_{GM} t/N).$$

Next, we consider Brownian particles in a convex or concave harmonic potential, $U(\vec{x}) = \theta\vec{x}^2$, described by the Orstein-Uhlenbeck process. The probability distribution given an initial position y is

$$p(x \mid y; t) = \prod_{\alpha=1}^{d} \frac{\exp\{-\theta(x_\alpha - y_\alpha \mathbf{e}^{-\theta t})^2/[2D_\alpha(1-\mathbf{e}^{-2\theta t})]\}}{\sqrt{2\theta^{-1}\pi D_\alpha(1-\mathbf{e}^{-2\theta t})}}$$

meaning that $D_\alpha(t) = 2D_\alpha\theta^{-1}(1-\mathbf{e}^{-2\theta t})$ and $\lambda_\alpha(t) = e^{-\theta t}$. Thus,

$$S_R = \frac{1}{2}\log\left[(2\pi e)^d \prod_{\alpha=1}^{d} \frac{\sigma_\alpha^2 D_\alpha(1-\mathbf{e}^{-2\theta t})}{\sigma_\alpha^2 N\theta \mathbf{e}^{-2\theta t} + D_\alpha(1-\mathbf{e}^{-2\theta t})}\right].$$

In the limit of infinite σ's, we get two very different long-time behaviors depending on the sign of θ. For $\theta > 0$ we have a harmonic trap. As $t \to \infty$, $S_R \sim d\theta t$. For $\theta < 0$, we have a potential that tends to quickly force particles away from the origin. In this case,

$$S_R = (d/2)\log\left[(2\pi \mathbf{e})(1-\mathbf{e}^{2|\theta|t})D_{GM}/(N|\theta|)\right].$$

Therefore, as $t \to \infty$, $S_R \sim \text{const.} - \frac{d}{2}e^{-2|\theta|t}$. Thus, after some initial transient loss of information, our ability to reconstruct the initial state plateaus, i.e., the system always retains information about its initial state for arbitrarily long times [see Fig. 1(a)]. For finite σ's, S_R has three distinct temporal regimes. It starts logarithmic, crosses over to linear, and then finally saturates to S_0 [see Figs. 1(b) and 2].

In Fig. 1(a), we have plotted the average retrodiction entropy as a function of time for five particles in potentials with various concavities [θ parameters, $U(x) = \theta x^2$]. The prior is a non-normalizable uniform prior. The process is an Ornstein-Uhlenbeck process when $\theta \neq 0$, and is the Weiner process when $\theta = 0$. For concave potentials (in blue), the retrodiction entropy converges to a finite value. For a potential with $\theta = 0$, we recover the Wiener process, and S_R increases logarithmically. For convex potentials, S_R is asymptotically linear, diverging much more quickly than the Weiner process.

In Fig. 1(b), we have shown the analogous plot, but for a Gaussian prior. For concave potentials, the retrodiction entropy still saturates to a value below the prior entropy value. For convex potentials, the retrodiction entropy starts logarithmic, becomes linear, and then quickly saturates to S_0. For the Wiener process, the retrodiction entropy does eventually approach the value of S_0, though very slowly–at $t = 1000$, it is still 2.5% away from S_0.

In Fig. 2(a), we show the time dependence of the entropies S_0, $\langle S_T\rangle$, S_t, and $\langle S_R\rangle$ for two particles in a convex potential

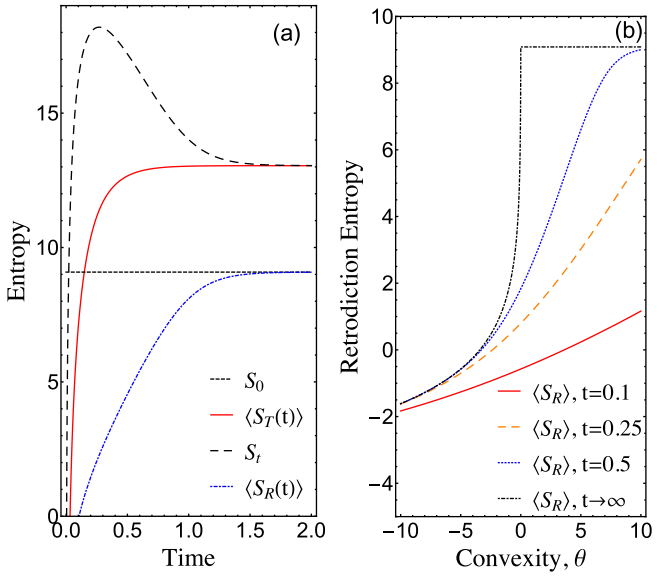

FIG. 2. Role of convexity. The lines for potential and the labels both go from top to bottom, with the exception of the dashed line for S_0. Left: Prior entropy S_0, average thermodynamic entropy $\langle S_T\rangle$, and observational entropy S_t for two particles in a harmonic trap, as derived in Appendix A. The retrodiction entropy is related to the other three, through $\langle S_R\rangle = \langle S_T\rangle - (S_t - S_0)$. The prior distribution is Gaussian with $\sigma = 5$. Right: S_R for the Ornstein-Uhlenbeck process at different times with a Gaussian prior, $\sigma = 5$. For positive convexity, S_R converges to S_0, for negative convexity, S_R converges to some smaller value, meaning some information can still be recovered.

with a Gaussian prior. This illustrates the fact that $\langle S_R\rangle = \langle S_T\rangle - (S_t - S_0)$. The linear behavior of S_R in the intermediate regime can be seen before it exponentially approaches the value of the entropy of the prior, S_0.

In Fig. 2(b), we plot the average retrodiction entropy of the Ornstein-Uhlenbeck process at specific times, starting with a Gaussian prior. In the long time limit, if the convexity is positive, the retrodiction entropy approaches the entropy of the prior distribution, S_0, and hence the black line being flat for all $\theta \geqslant 0$. However, if the convexity is negative (so a concave potential), we can see that the retrodiction entropy converges to a value less than S_0. This indicates that by making a measurement, we gain information about the initial state of the system even after arbitrarily long times.

V. RETRODICTION OF A CHAOTIC SYSTEM

To study how chaos relates to retrodictability we consider the simplest of chaotic systems, the logistic map,

$$X_t = r \cdot X_{t-1}(1 - X_{t-1}),$$

characterized by a single parameter r, which determines whether the system is chaotic. Our key result here is somewhat counterintuitive: We find that the system is maximally retrodictable right before and right after it transitions into chaos.

The asymptotic properties of the logistic map is well known [22]. The values p_n take as n tends to infinity, i.e., the attractors, is shown in the bifurcation diagram [Fig. 3(a)]. For small values of r, the trajectories are periodic. As r is increased, there is a

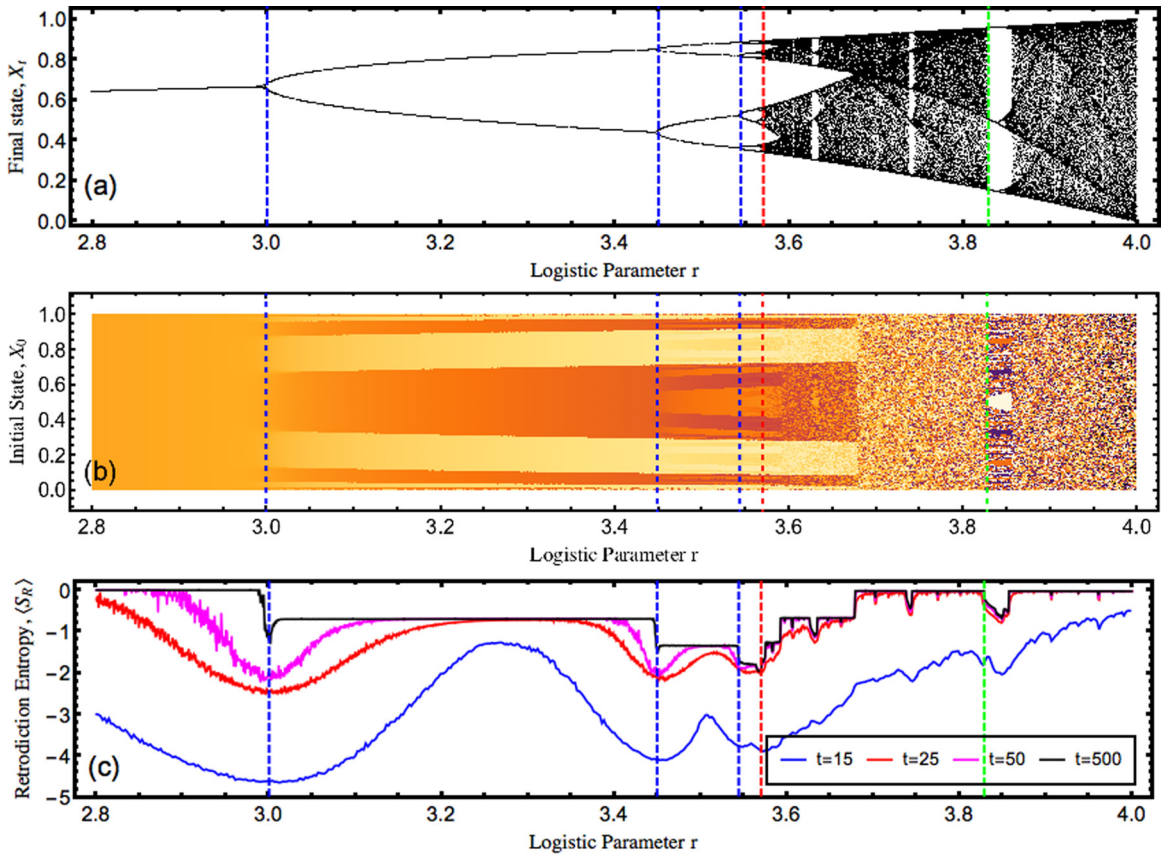

FIG. 3. Retrodiction entropy, bifurcations, and chaos. The vertical dashed lines from left to right are (i) period doubling, (ii) period quadrupling, (iii) period $\times 8$, (iv) onset of chaos (red), and (v) the onset of one particular island of stability where chaos breaks off to periodic motion (green). Top: The bifurcation diagram for the logistic map, showing X_t for multiple large t values. Middle: Basins of attraction. The initial state X_0 determines the final state X_t, ($t = 200$) within [0,1], which is mapped to a color gradient from dark red (0) to light yellow (1). The change in the number of basins can be clearly seen near the vertical lines. As the system transitions into chaos, nearby points start converging to distinct final points. The system becomes chaotic (at $r \simeq 3.58$) and then well mixed (at $r \simeq 3.68$. Bottom: The (normalized) retrodiction entropy vs logistic parameter r is plotted at several different times, t. Lines from top to bottom are $t = 500$, $t = 50$, $t = 25$, $t = 15$. The $t = 500$ line is an excellent approximation to the asymptotic limit of $\langle S_R \rangle$. Note that in the non-chaotic regime, the retrodiction entropy converges to flat steps, whereas in the chaotic regime, the retrodiction entropy converges to steps (with values equal to that in the nonchaotic regime) with occasional dips coinciding with islands of stability. Course graining was done with $b = 500$ bins, with 10000 sample initial points per bin.

sequence of period doublings (cf. Fig. 3, blue vertical dashes) until the system transitions to chaos at $r \simeq 3.57$ (red vertical dashes). Within the chaotic regime, there are occasional islands of stability where periodic attractors exist. For example, at $r \simeq 3.83$ there is a period three attractor (green vertical dashes).

Since the logistic map is purely deterministic, in order to define probabilities and entropies we suppose that the state of the system cannot be measured with infinite accuracy—similar to how probability and entropy arise in classical statistical mechanics. To avoid artifacts stemming from the precise details of coarse graining, we pick very small bins with randomized positions

Specifically, we coarse grain the interval [0,1] randomly into b bins by picking $b-1$ random numbers uniformly and ordering them $0 < x_1 < x_2 < \ldots < x_{b-1} < 1$. We then uniformly and randomly sample s points from each bin, and iterate each point τ times via the logistic map. This way, we construct the probability transition matrix $T_{ji}^{(\tau)}$, the probability that a point selected randomly from bin j ends in bin i after τ logistic steps. Using this, and assuming a uniform prior on picking the initial point, we can obtain the retrodiction probability matrix $R_{ji}^{(\tau)}$, and the average retrodiction entropy.

As the binning is random, the value of average retrodiction entropy is slightly different for each realization of the binning, so we average over many different random binnings. We note that we are essentially calculating the information dimension of the retrodiction probability. Information dimension [23,24] is one of several common ways to calculate fractal dimension. Our prescription here is only different in that we are applying it to our retrodiction of the original state, not to the calculation of the final state.

Figure 3 contains several panels related to the retrodictability of the logistic map. The top panel is the bifurcation diagram for the logistic map, which we align with the other two panels to use as a reference.

The middle panel shows what initial states converge to what final state. Here we see the basins of attraction of the logistic map. The vertical axis indicates the initial position of

the point, whereas the color represents the value the point has after 250 iterations. We can see how the unit interval splits into domains at each bifurcation point. At the onset of chaos, even the points very near each other can end up in different phase oscillations. The degree of chaos increases several times, when subdomains of the unit interval become more mixed. This occurs for example at $r = 3.58$ and $r = 3.59$ before the point of complete mixing at $r = 3.68$.

The bottom panel shows the retrodiction entropy at various times. The black line, for $t = 500$ steps, is a good approximation of the asymptotic limit of $\langle S_R \rangle$. For parameter values below the first period doubling, retrodiction entropy is at a maximum since all points in the unit interval converge to a single value, therefore observing that value does not provide any useful information about the initial state of the system. Therefore, $S = \log V = 0$ since the volume V is the unit interval. At the period doubling, the asymptotic value of S_R drops to $-\log 2$. This is reflective of the fact that in the two period region, the measure of the set of points that converge to each period is $1/2$. Therefore, the retrodiction entropy given either of the two ending positions is $-\log 2$. This trend of reduction in average retrodiction entropy continues with every period doubling, as an equal measure of points converge to different basins.

Note that, as period doublings occur more rapidly with increasing r, our finite bin size prohibits us from resolving the discrete steps close to the onset of chaos. As period doublings happen exponentially quickly and exponentially close together, an exponential number of bins becomes necessary to distinguish between the entropy drops associated with successive bifurcations.

The blue vertical dashed lines in Fig. 3 show the locations of the period doublings. Near the period doubling points, there is a dramatic slowdown in convergence of S_R to its asymptotic value, which is reflective of the fact that there is a slowdown in convergence of sequences to the periodic attractor.

As period multiplicities of every power of 2 occur before the onset of chaos, the long-time limit of differential retrodiction entropy approaches negative infinity (in the limit of infinite number of bins). Even with a limited number of bins, the asymptotic retrodiction entropy hits a minimum right at the chaotic transition.

Past the point of chaos, retrodiction entropy ascends in steps with the same asymptotic values as the descending steps. The reason why the steps have the same value can be seen in the middle panel of Fig. 3. As r approaches chaos, the system breaks the unit interval of starting positions into subdomains that map to different periodic attractors (which are subdivided somewhat similarly to a Cantor set). After the onset of chaos, the subdomains undergo mixing, as previously mentioned, where any point that started in that domain has an equal chance of ending up in any attractor in any subdomain of that domain.

The reconstruction entropy in the chaotic regime also has occasional dips, which correlate with the islands of stability. For example, we have marked the value $r = 3.83$ in green, which is where the logistic map has a period three oscillation. The dips around $r = 3.63$ and $r = 3.74$ occur because the logistic map is not chaotic for some values of (x, r), but instead an entire neighborhood in the unit interval converges to the same attractor.

VI. DISCUSSION

The approach of using retrodiction entropy bears some similarities to other methods of inference, particularly maximum *a posteriori* (MAP) estimation and other Bayesian methods, but also has significant differences. Philosophically, our goal in defining S_R is not to find the mode of a distribution (this is the usual goal of Bayesian inference), but to characterize the information contained in the distribution as a whole. Identifying modes, or the most likely initial state can be very misleading. For example, in highly degenerate systems, there could be many peaks in $\mathcal{R}$, each containing a small amount of probability mass. In contrast, S_R characterizes the information content within the entire probability distribution.

That being said, entropy does not constitute a complete characterization of a probability distribution either. For example, it might be informative to pull out a guess from $\mathcal{R}$ and compare it with the actual initial state,

$$\int \mathcal{R}_y(x_1)(x_1 - x_2)^2 \mathcal{R}_y(x_2) \mathbf{d}x_1 \mathbf{d}x_2.$$

Since entropy does not take into account information about the spatial location of probability mass, it would not inform on this quantity.

A. Comparison with other approaches

There is a long history of inference and information theory in the development of statistical mechanics. Here, we briefly review a few similar methods of doing inference and measuring predictability.

Problems in inverse statistical mechanics are generally solved by using maximum likelihood estimation (MLE) or, if prior information is available, maximum *a posteriori* (MAP) estimation. Other methods are available, for example, the pseudolikelihood [25]. However, most of the problems typically treated in inverse statistical physics are lattice problems, and the typical goal is to find microscopic parameters of the system given some number of (generally independent) measurements, rather than finding the state of the system in the past. For example, a prototypical inverse statistical mechanical problem is the inverse Ising problem [26], where the connections J_{ij} between spin variables is unknown, the spin configuration is sampled some number of times from the equilibrium distribution, and the problem is to infer the most likely matrix J_{ij}.

A series of papers by Crutchfield and Ellison treat semi-infinite chains of random variables as consecutive states in discrete time, and suggests that the mutual information between semi-infinite sets of variables is a good measure for the amount of information about the past stored in the present [27–30]. Their backwards entropy $h_\mu = \lim_{n\to\infty} H(X_{-n+1}, \ldots, X_0)/n$ differs from our retrodiction entropy, which, in compatible notation, becomes $\langle S_R \rangle = H(X_0) - I(X_0; X_t)$ [cf. (8)]. Note that while h_μ is defined for a chain of infinite time points, retrodiction entropy operates between two specific times.

The goals of computational mechanics and our retrodiction entropy approach are different. Computational mechanics asks what finite-state machine can statistically reproduce a sequence or random variables. Furthermore, many of the examples they treat are not physical systems, but finite-state computational

processes, they look at, e.g., the random insertion process [27], random noisy copy, and the golden mean process [28], though in Ref. [30] the authors look at reproducing the patterns in different Ising systems.

In addition, the constraint of having infinite pasts and futures amounts to studying systems only in equilibrium, which is not a case we would typically be interested in when studying retrodiction entropy.

B. Possible generalizations

We can loosen our formalism to make it applicable to general inference problems; not just problems in statistical mechanics. An inference problem is typically of the form where there is a space of sets of possible model parameters, A, and a space of possible observed outcomes, Ω. The transition probability is the probability that an observable event occurs given a set of model parameters. There is not necessarily any variable that serves as time. As the problem is one of reconstructing parameters, and there is no time, so no past, we would call the Bayesian inverse of $\mathcal{T}$ reconstruction probability and call the corresponding S_R reconstruction entropy (instead of retrodiction probability and entropy).

Reconstruction entropy is a measurement of how well we can determine the parameters of a system given an observed event generated from a model with unknown parameters. Retrodiction entropy is a special case of this where the set of parameters is the same as the set of observables ($A = \Omega$), e.g., both are phase space. Additionally, when retrodicting, we consider a parameterized family of transition probabilities, understanding this parameter to be our system time. For the more general reconstruction entropy, most of the formulas we have derived still hold, for example Eqs. (1)–(4), (6), and (8), and the KL divergence relations in Appendix B. On the other hand, results such as (7) do not hold if there is no time parameter.

VII. CONCLUSION

We introduced the notion of retrodiction entropy as a measure of our ability to infer the past state of a collection of particles based on a single measurement of the system, and derived a relationship between this and thermodynamic entropy. We have established bounds on the retrodiction entropy generation rate, derived a set or KL divergence relations between different relevant probabilities, and outlined retrodiction entropy's asymptotic properties. We also showed that for systems where the initial state is an equilibrium distribution, the average forward and retrodiction entropy are identical. Lastly, we analytically solved two concrete examples, quantifying how rapidly a system of particles forgets its initial state in convex, concave, and flat potentials, and analyzing macrostate retrodiction entropy for a chaotic system. Particularly, we saw that in a concave potential there is an upper limit to the loss of information pertaining the initial state, and for the logistic map, we saw sharp changes in asymptotic retrodiction entropy at period doublings, and could identify islands of stability in the chaotic regime by dips in retrodiction entropy.

The connection between thermodynamic quantities $\langle S_T \rangle$, S_t and a purely information theoretical one, S_R, is in accordance with the seminal works of Maxwell, Smoluchowski, Landauer, Szillard, Beckenstein, and others [1–4]. We now know, from (3), that thermodynamic entropy at present time not only quantifies the information content of the state of the system at present time, it also relates to how precisely information about the original state of the system can be recovered after some amount of time has passed.

APPENDIX A: DERIVATION OF THE RELATIONSHIP BETWEEN RETRODICTION ENTROPY AND THERMODYNAMIC ENTROPY

We use sum notation throughout, although these could be replaced with integrals. Suppose P is normalized. Then (3) can be proved through simple integration:

$$\begin{aligned}\langle S_R \rangle &= \sum_\omega P_t(\omega) S_R(\xi) = -\sum_\omega P_t(\omega) \sum_\alpha \mathcal{R}_\omega(\alpha) \log \mathcal{R}_\omega(\alpha) \\ &= -\sum_{\omega,\alpha} \mathcal{T}_\alpha(\omega) P_0(\alpha) \log\left(\frac{\mathcal{T}_\alpha(\omega) P_0(\alpha)}{P_t(\omega)}\right) \\ &= -\sum_{\omega,\alpha} P_0(\alpha) \mathcal{T}_\alpha(\omega) \log \mathcal{T}_\alpha(\omega) + \sum_\omega P_t(\omega) \log P_t(\omega) \\ &\quad - \sum_\alpha P_0(\alpha) \log P_0(\alpha) = \langle S_T \rangle - (S_t - S_0).\end{aligned}$$

where we substituted $P_t(\omega) = \sum_\alpha \mathcal{T}_\alpha(\omega) P_0(\alpha)$.

As an explicit example of this, consider the Gaussian process family we discussed in the paper, with $\mathcal{T}, P_0, P_t$ given by (11), (12), and (13). For this case,

$$\begin{aligned}S_T &= \langle S_T \rangle = \frac{1}{2} \log\left[\pi^N \prod_{\alpha=1}^d D_\alpha(t)^N\right] + \frac{Nd}{2} \\ S_0 &= \frac{1}{2} \log\left[(2\pi e)^d \prod_{\alpha=1}^d \sigma_\alpha^2\right] \\ S_t &= \frac{1}{2} \log\left[(2\pi^N N)^d \prod_{\alpha=1}^d \sigma_\alpha^2 D_\alpha(t)^N \frac{\lambda_\alpha(t)^2}{D_\alpha(t)\kappa_\alpha(t)}\right] + \frac{Nd}{2}\end{aligned}$$

from which it can be shown, using (3), that

$$S_R = \langle S_R \rangle = \frac{1}{2} \log\left[\left(\frac{\pi e}{N}\right)^d \prod_{\alpha=1}^d \frac{D_\alpha(t)\kappa_\alpha(t)}{\lambda_\alpha(t)^2}\right].$$

APPENDIX B: KL-DIVERGENCE RELATIONS

Here, we derive (6). We start with the definition of KL divergence:

$$\begin{aligned}&D[\mathcal{R}_{\omega_1} \| \mathcal{R}_{\omega_2}] \\ &= -\sum_\alpha \mathcal{R}_{\omega_1}(\alpha) \log \frac{\mathcal{R}_{\omega_2}(\alpha)}{\mathcal{R}_{\omega_1}(\alpha)} \\ &= -\sum_\xi \frac{\mathcal{T}_\alpha(\omega_1) P_0(\alpha)}{P_t(\omega_1)} \left[\log\left(\frac{\mathcal{T}_\alpha(\omega_2)}{\mathcal{T}_\alpha(\omega_1)}\right) + \log\left(\frac{P_t(\omega_1)}{P_t(\omega_2)}\right)\right].\end{aligned}$$

Averaging over ω's with the probability weight $P_t(\omega_1)P_t(\omega_2)$, the first term in the brackets gives

$$\begin{aligned}
&-\sum_{\omega_1,\omega_2,\alpha} P_t(\omega_2)\mathcal{T}_\alpha(\omega_1)P_0(\alpha)\log[\mathcal{T}_\alpha(\omega_2)/\mathcal{T}_\alpha(\omega_1)]\\
&=-\sum_{\omega_1,\omega_2,\alpha} P_0(\alpha)P_t(\omega_2)[\mathcal{T}_\alpha(\omega_1)\log\mathcal{T}_\alpha(\omega_2)\\
&\quad-\mathcal{T}_\alpha(\omega_1)\log\mathcal{T}_\alpha(\omega_1)]\\
&=-\sum_{\omega_1,\omega_2,\alpha} P_0(\alpha)P_t(\omega_2)\log\mathcal{T}_\alpha(\omega_2)\\
&\quad-P_0(\alpha)\mathcal{T}_\alpha(\omega_1)\log\mathcal{T}_\alpha(\omega_1)\\
&=-\langle S_T\rangle-\sum_{\omega_2} P_t(\omega_1)\log\mathcal{T}_\alpha(\omega_2)\\
&=S_t-\langle S_T\rangle+\sum_\alpha D[P_t\|\mathcal{T}_\alpha]
\end{aligned}$$

(we have used the fact that $D[A\|B]=-\sum A\log B-S_A$ and $\sum_\omega \mathcal{T}_\alpha(\omega)=1$) whereas the second term gives

$$\begin{aligned}
&-\sum_{\omega_1,\omega_2,\alpha} P_t(\omega_2)\mathcal{T}_\alpha(\omega_1)P_0(\alpha)\log\frac{P_t(\omega_1)}{P_t(\omega_2)}\\
&=-\sum_{\omega_1,\omega_2} P_t(\omega_2)\left(\sum_\alpha \mathcal{T}_\alpha(\omega_1)P_0(\alpha)\right)\log\frac{P_t(\omega_1)}{P_t(\omega_2)}\\
&=-\sum_{\omega_1,\omega_2} P_t(\omega_2)P_t(\omega_1)\log\frac{P_t(\omega_1)}{P_t(\omega_2)}\\
&=S_t-S_t=0.
\end{aligned}$$

Putting everything together,

$$\langle D[\mathcal{R}_{\omega_1}\|\mathcal{R}_{\omega_2}]\rangle=\langle D[\mathcal{N}\|\mathcal{T}_\xi]\rangle+S_t-\langle S_T\rangle.$$

The second term here is,

$$\begin{aligned}
\langle D[\mathcal{N}\|\mathcal{T}_\xi]\rangle&=-\sum_{\xi,\omega} P(\xi)\mathcal{N}(\omega)\log\frac{\mathcal{T}_\xi(\omega)}{\mathcal{N}(\omega)}\\
&=-\sum_{\xi_1,\xi_2} P(\xi_1)P(\xi_2)\mathcal{T}_{\xi_1}(\omega)\log\mathcal{T}_{\xi_2}(\omega)-S_t\\
&=\langle D[\mathcal{T}_{\xi_1}\|\mathcal{T}_{\xi_2}]\rangle+\langle S_T\rangle-S_t.
\end{aligned}$$

Putting these equations together gives us Eq. (6).

We can take the KL divergence between any pair of distributions that have a common domain. It is natural to only compare distributions that are either both on the final state or both on the initial state. Furthermore, as the KL divergence is asymmetric, we can ask about both orderings. The six options are $(\mathcal{T},\mathcal{T})$, $(\mathcal{T},P_t)$, $(P_t,\mathcal{T})$, $(\mathcal{R},\mathcal{R})$, $(\mathcal{R},P_0)$, and $(P_0,\mathcal{R})$. In a similar way to our derivations above, we can find relations between the averages of the KL divergence between all these pair in terms of each other or in terms of entropies:

$$\begin{aligned}
\langle D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}]\rangle&=\langle D[P_0\|\mathcal{R}_\omega]\rangle+S_t-\langle S_T\rangle\\
\langle D[\mathcal{T}_\alpha\|P_t]\rangle&=\langle D[\mathcal{R}_\omega\|P_0]\rangle=S_0-\langle S_R\rangle=S_t-\langle S_T\rangle\\
\langle D[P_t\|\mathcal{T}_\alpha]\rangle&=\langle D[P_0\|\mathcal{R}_\omega]\rangle\\
\langle D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}]\rangle&=\langle D[\mathcal{R}_{\omega_1}\|\mathcal{R}_{\omega_2}]\rangle.
\end{aligned}$$

One can put these together to derive relations for the averages of the symmetric combinations of KL divergences.

$$\begin{aligned}
\langle D[\mathcal{T}_\alpha\|P_t]+D[P_t\|\mathcal{T}_\alpha]\rangle&=\langle D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}]\rangle\\
\langle D[\mathcal{R}_\omega\|P_0]+D[P_0\|\mathcal{R}_\omega]\rangle&=\langle D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}]\rangle.
\end{aligned}$$

APPENDIX C: LIMITS ON THE SIZE OF OBSERVATIONAL AND RETRODICTION ENTROPY

We can use Jensen's inequality to put an upper bound on the time rate of change of S_t. Since $-\log x$ is a convex function, we have the inequality,

$$-\log\sum_\omega P_0(\alpha)\mathcal{T}_\alpha(\omega)\leqslant-\sum_\alpha P_0(\alpha)\log\mathcal{T}_\alpha(\omega).$$

Start with the definition of S_t, then apply Jensen's inequality:

$$\begin{aligned}
\dot{S}_t&=-\sum_\xi \dot{P}_t(\omega)\log P_t(\omega)=-\sum_\omega \dot{P}_t(\omega)\log\sum_\alpha P_0(\alpha)\mathcal{T}_\alpha(\omega)\\
&\leqslant-\sum_{\alpha,\omega} P_0(\alpha)\dot{P}_t(\omega)\log\mathcal{T}_\alpha(\omega)\\
&=-\sum_{\alpha,\omega} P_0(\alpha)\dot{P}_t(\omega)\left(\log\frac{\mathcal{T}_\alpha(\omega)}{P_t(\omega)}+\log P_t(\omega)\right)\\
&=-\sum_{\alpha,\omega} P_0(\alpha)\dot{P}_t(\omega)\log\frac{\mathcal{T}_\alpha(\omega)}{P_t(\omega)}+\dot{S}_t.
\end{aligned}$$

Canceling the $\dot{S}_N$ terms on both sides yields

$$0\leqslant-\sum_\alpha P_0(\alpha)\sum_\omega \dot{P}_t(\omega)\log\frac{\mathcal{T}_\alpha(\omega)}{P_t(\omega)},$$

which bears some similarity to the KL divergence. The derivative of an arbitrary KL divergence is

$$\frac{\partial}{\partial t}D[p\|q]=-\sum\dot{p}\log\frac{q}{p}-\sum\frac{p}{q}\dot{q}.$$

Using this in the preceding inequality, we get

$$\begin{aligned}
0&\leqslant\frac{\partial}{\partial t}\langle D[P_t\|\mathcal{T}_\alpha]\rangle+\sum_\alpha P_0(\alpha)\sum_\omega P_t(\omega)\frac{\partial}{\partial t}\log\mathcal{T}_\alpha(\omega)\\
&=\frac{\partial}{\partial t}\langle D[P_t\|\mathcal{T}_\alpha]\rangle+\sum_\alpha P_0(\alpha)\sum_\omega P_t(\omega)\frac{\partial}{\partial t}\log\frac{\mathcal{R}_\omega(\alpha)P_t(\omega)}{P_0(\alpha)}\\
&=\frac{\partial}{\partial t}\langle D[P_t\|\mathcal{T}_\alpha]\rangle+\sum_\omega P_t(\omega)\sum_\alpha P_0(\alpha)\frac{\partial}{\partial t}\log\frac{\mathcal{R}_\omega(\alpha)}{P_0(\alpha)}\\
&=\frac{\partial}{\partial t}\langle D[P_t\|\mathcal{T}_\alpha]\rangle-\left\langle\frac{\partial}{\partial t}D[P_0\|\mathcal{R}_\omega]\right\rangle.
\end{aligned}$$

Using the expression we previously discussed for $\langle D[P_t\|\mathcal{T}_\alpha]\rangle$, we can reintroduce $\dot{S}_t$ to the equation,

$$\dot{S}_t\leqslant\langle\dot{S}_T\rangle+\frac{\partial}{\partial t}\langle D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}]\rangle-\left\langle\frac{\partial}{\partial t}D[P_0\|\mathcal{R}_\omega]\right\rangle.$$

We can also write this as a lower bound on $\frac{\partial}{\partial t}\langle S_R\rangle$ via (3)

$$\frac{\partial}{\partial t}\langle S_R\rangle\geqslant-\frac{\partial}{\partial t}\langle D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}]\rangle+\left\langle\frac{\partial}{\partial t}D[P_0\|\mathcal{R}_\omega]\right\rangle.\tag{C1}$$

Now we will make use of the fact that for a Markov process, the relative entropy of two distributions is nonincreasing [21]. We include this theorem below for the sake of completeness.

Theorem. Consider two probability distributions p, q, on the same state space. Then at any times $t_1 < t_2$,

$$D[p_{t_1}\|q_{t_1}] \geqslant D[p_{t_2}\|q_{t_2}].$$

Proof. Let $s < t$. Then,

$$\begin{aligned} &D[p(x_t|x_s)\|q(x_t|x_s)] \\ &= D[p(x_t)\|q(x_t)] + D[p(x_s|x_t)\|q(x_s|x_t)] \\ &= D[p(x_s)\|q(x_s)] + D[p(x_t|x_s)\|q(x_t|x_s)]. \end{aligned}$$

By the definition of Markov, $p(x_t|x_s) = q(x_t|x_s)$, so $D[p(x_t|x_s)\|q(x_t|x_s)] = 0$. Then, subtracting the second and third lines, we get

$$D[p_t\|q_t] - D[p_s\|q_s] = -D[p_{s,t}\|q_{s,t}] \leqslant 0. \quad \blacksquare$$

If our forward dynamics are Markovian (as they are, for example, in the case of diffusion), this theorem holds and $\frac{\partial}{\partial t} D[\mathcal{T}_{\alpha_1}\|\mathcal{T}_{\alpha_2}] \leqslant 0$ for all α_1, α_2. Therefore, the first term on the right-hand side of Eq. (C1) is non-negative.

The second term of Eq. (C1) is harder to work with. Intuitively, we expect $\mathcal{R}$ to approach P as we lose information about the past due to stochastic events. So we expect $D[P\|\mathcal{R}_\xi]$ to eventually reach a minimum for any fixed ξ. As long as $\langle D[P\|\mathcal{R}_\xi]\rangle$ decreases more slowly than $\langle D[\mathcal{T}_{\omega_1}\|\mathcal{T}_{\omega_2}]$, this bound is good enough to guarantee that $\partial\langle S_R\rangle/\partial t \geqslant 0$.

[1] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[2] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[3] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Champaign, 1998).
[4] H. S. Leff and A. F. Rex, *Maxwell's Demon: Entropy, Information, Computing* (Princeton University Press, Princeton, 2014).
[5] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Vol. 40 (John Wiley & Sons, New York, 2011).
[6] M. Welling and Y. W. Teh, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 681–688.
[7] B. A. Desmarais and S. J. Cranmer, Physica A **391**, 1865 (2012).
[8] V. A. T. Nguyen and D. C. Vural, Phys. Rev. E **96**, 032314 (2017).
[9] P. C. Hansen, J. G. Nagy, and D. P. O'leary, *Deblurring Images: Matrices, Spectra, and Filtering* (SIAM, Philadelphia, 2006).
[10] R. H. Chan and K. Chen, SIAM J. Sci. Comput. **32**, 1043 (2010).
[11] P. Ullersma, Physica **32**, 27 (1966).
[12] H.-Y. Yu, D. M. Eckmann, P. S. Ayyaswamy, and R. Radhakrishnan, Phys. Rev. E **91**, 052303 (2015).
[13] W. T. Coffey and Y. P. Kalmykov, *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry and Electrical Engineering*, Vol. 27 (World Scientific, Singapore, 2012).
[14] F. Wolf, J. Math. Phys. **29**, 305 (1988).
[15] M. Hashemi, Physica A **417**, 141 (2015).
[16] M. Bernstein and L. S. Brown, Phys. Rev. Lett. **52**, 1933 (1984).
[17] J. A. Carrillo and G. Toscani, Math. Meth. Appl. Sci. **21**, 1269 (1998).
[18] G. Toscani, Q. Appl. Math. **57**, 521 (1999).
[19] V. Schwämmle, E. M. Curado, and F. D. Nobre, Europhys. J. B **58**, 159 (2007).
[20] A. R. Plastino, H. G. Miller, and A. Plastino, Phys. Rev. E **56**, 3927 (1997).
[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley and Sons, New York, 2012).
[22] R. M. May, Nature (London) **261**, 459 (1976).
[23] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[24] J. D. Farmer, Zeitschrift für Naturforschung A **37**, 1304 (1982).
[25] J. Besag, J. Roy. Stat. Soc. B 192 (1974).
[26] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).
[27] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney, Phys. Rev. Lett. **103**, 094101 (2009).
[28] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield, J. Stat. Phys. **136**, 1005 (2009).
[29] J. P. Crutchfield and C. J. Ellison, arXiv:1012.0356.
[30] D. P. Feldman and J. P. Crutchfield (unpublished), https://www.santafe.edu/research/results/working-papers/discovering-noncritical-organization-statistical-m.