COMMUNICATIONS PHYSICS

ARTICLE

https://doi.org/10.1038/s42005-019-0159-z OPEN

Enhancing the predictability and retrodictability of stochastic processes

Nathaniel Rupprecht [1] & Dervis Can Vural [1]

Scientific inference involves obtaining the unknown properties or behavior of a system in the light of what is known, typically without changing the system. Here we propose an alternative to this approach: a system can be modified in a targeted way, preferably by a small amount, so that its properties and behavior can be inferred more successfully. For the sake of concreteness we focus on inferring the future and past of Markov processes and illustrate our method on two classes of processes: diffusion on random spatial networks, and thermalizing quantum systems.

[1] University of Notre Dame, Main Building, Notre Dame, IN, USA. Correspondence and requests for materials should be addressed to D.C.V. (email: dvural@nd.edu)

Much of science revolves around inference, reconstructing the unknown from what is known[1–3]. Observable patterns here and now inform us of inaccessible patterns out and away. For example, using inferential techniques, one can reconstruct the history of life from available fossils[4–6], or predict the fate of the universe by observing the present night sky[7–9]; one can infer hidden states and transition probabilities[10–12], connections and weights of neural networks[13–15] or parameters, initial states and interaction structures of complex systems[16–25].

Ordinarily, inference is a passive, non-disruptive process. Unlike engineering, natural sciences are motivated by knowing nature, rather than changing it. However, knowing and changing are not necessarily mutually exclusive. Earlier, it was established that attempting to describe and predict a system can inadvertently influence it, potentially even rendering it indescribable and unpredictable[26,27]. Here we study the converse case of how the intrinsic properties of a system can be purposefully modified so that its past or future is more inferable.

A number of authors addressed the problem of predicting the future and retrodicting the past of a stochastic process[28–31]. In this study, we are concerned not with finding strategies or algorithms to predict or retrodict stochastic systems, but rather with optimally modifying systems so that their predictability or retrodictability increases.

An engineer might use control theory to balance a bipedal robot, stabilize the turbulent flows surrounding a wing, or maximize the signal to noise ratio in an electric circuit[32]. Here we do the same, but optimize the susceptibility of a system to the inquiry of its past and future.

Forward in time, the entropy associated with the probability distribution of the system state will increase monotonically, as per H-theorems[33–35]. A similar trend also holds backwards in time[31]. Here we determine how transition rates should be perturbed infinitesimally as to minimize the generation of inferential entropy in either temporal direction. After establishing a general theoretical framework, we implement these ideas to two specific example systems. The first is a diffusion process taking place on a spatial random network. The second is a quantum harmonic oscillator with a time-dependent temperature.

Results

Quantifying predictability and retrodictability. The past and future of a stochastic system with a concentrated and sharply peaked probability distribution can be inferred with high certainty. Accordingly, we use the Gibbs–Shannon entropy to quantify the inferrability of a system[36], and later on show that this indeed is a good measure. Given a stochastic process, X_t, characterized by its transition matrix $T_\alpha(\omega) = Pr(X_t = \omega | X_0 = \alpha)$, and initial state α, the entropy of the process at a final time t is

$$S_{\mathrm{T}}(\alpha) = -\sum_{\omega} T_\alpha(\omega) \log T_\alpha(\omega). \tag{1}$$

When X_t is the state of a thermodynamic system, this is the standard thermodynamic entropy. In the present information-theoretical context, we refer to S_{T} as the "prediction entropy".

Naturally, the average entropy generated by a process depends on how it is initialized—the prior distribution $P^{(0)}$. To characterize the the process itself, we marginalize over the initial state, α, $\langle S_{\mathrm{T}} \rangle = \sum_{\alpha} P^{(0)}(\alpha) S_{\mathrm{T}}(\alpha)$, where $P^{(0)}(\alpha)$ is the probability of starting at α. Likewise, we quantify the retrodictability of a process by a "retrodiction entropy", $\langle S_{\mathrm{R}} \rangle = \sum_{\omega} P^{(t)}(\omega) S_{\mathrm{R}}(\omega)$. Here, $R_\omega(\alpha)$ is the probability the system started in state α given that the observed final state was ω, S_{R} is its entropy analogous to Eq. (1), and $P^{(t)}(\omega)$ is the probability that the process is in state ω at time t unconditioned on its initial state.

Interestingly, the predictability and retrodictability of a system are tightly connected: Since S_{T} and S_{R} are related by the Bayes' theorem, $R_\omega(\alpha) = T_\alpha(\omega) P^{(0)}(\alpha) / \sum_{\alpha'} T_{\alpha'}(\omega) P^{(0)}(\alpha')$, it follows that $\langle S_{\mathrm{T}} \rangle$ and $\langle S_{\mathrm{T}} \rangle$ are also related[31],

$$\langle S_{\mathrm{R}} \rangle = \langle S_{\mathrm{T}} \rangle - (S_t - S_0) \tag{2}$$

where S_0 is the entropy of the prior probability distribution $P^{(0)}$, and S_t is the entropy of $P^{(t)}(\omega) = \sum P^{(0)}(\alpha) T_\alpha(\omega)$.

We use $\langle S_{\mathrm{T}} \rangle$ and $\langle S_{\mathrm{R}} \rangle$ to measure how well we can predict the future and retrodict the past of a stochastic process. The higher the entropies, the less certain the inference will be.

Variations of Markov processes. In a Markov process, the state of a system fully determines its transition probability to other states. Markov processes accurately describe a number of phenomena ranging from molecular collisions through migrating species to epidemic spreads[37–41].

Consider such a Markov process defined by a transition matrix $\mathbf{T}$, with elements T_{ji}, which we will visualize as a weighted network. We assume that we know the transition rates and prior distribution over states at $t = 0$ with perfect accuracy, but do not know what state the system is in, except at the initial (final) time. From this data we will predict (retrodict) the final (initial) state of the system.

A system initialized in state i with probability $P_i^{(0)}$, upon evolving for t steps, will follow a new distribution $P_i^{(t)} = \sum_j P_j^{(0)} (T^t)_{ji}$. Accordingly,

$$\begin{aligned} \langle S_{\mathrm{R}} \rangle &= -\sum_{i,j} P_j^{(t)} (R^t)_{ji} \log (R^t)_{ji} \\ \langle S_{\mathrm{T}} \rangle &= -\sum_{i,j} P_j^{(0)} (T^t)_{ji} \log (T^t)_{ji}. \end{aligned} \tag{3}$$

Thus both entropies depend on the duration of the process t. Note that probability is normalized $\sum (T^t)_{ji} = 1$ for all j, t.

Suppose that it is somehow possible to change the physical parameters of a system slightly, so that the probability of transitions are perturbed, $T_{ji} \to T'_{ji} = T_{ji} + \epsilon q_{ji}$, where ϵ is a small parameter. For now, we do not assume any structure on q, other than implicitly demanding that it retains probabilities within [0, 1] and preserves the normalization of rows. This variation leads to a change in the t-step transition matrix,

$$\begin{aligned} (\mathbf{T} + \epsilon \mathbf{q})^t &= \mathbf{T}^t + \sum_{p=1}^{t} \epsilon^p \boldsymbol{\eta}^{(t,p)} \\ \boldsymbol{\eta}^{(t,p)} &= \sum_{1 \le k_1 < \ldots < k_p \le t} \mathbf{T}^{k_1 - 1} \boldsymbol{\xi}_{k_2 - k_1 - 1} \boldsymbol{\xi}_{k_3 - k_2 - 1} \cdots \boldsymbol{\xi}_{t - k_p} \end{aligned} \tag{4}$$

where $\boldsymbol{\xi}_k = \mathbf{q}\mathbf{T}^k$. The superscripts of $\boldsymbol{\eta}^{(t,p)}$ refer to the power of the transition matrix, t, and the order of the contribution, p, which is analogous to the order of the derivative of a function. So $\boldsymbol{\eta}^{(t,p)}$ is the p-th order contribution to the varied t-step transition matrix. This defines a set of p-th order effects for the nth power of the transition matrix. In the sequel, we will be studying first variations, therefore, we will only need

$$\boldsymbol{\eta}^{(t,1)} \equiv \boldsymbol{\eta}^{(t)} = \mathbf{q}\mathbf{T}^{t-1} + \mathbf{T}\mathbf{q}\mathbf{T}^{t-2} + \ldots + \mathbf{T}^{t-1}\mathbf{q}. \tag{5}$$

The difference between the the entropies of the perturbed and the original systems is $\Delta \langle S_{\mathrm{T,R}} \rangle \equiv \langle S_{\mathrm{T,R}}(\mathbf{T}') \rangle - \langle S_{\mathrm{T,R}}(\mathbf{T}) \rangle$. Whenever $\Delta \langle S_{\mathrm{T,R}} \rangle$ is of order ϵ and higher, we can evaluate the variation

$$\delta \langle S_{\mathrm{T,R}} \rangle = \lim_{\epsilon \to 0} \Delta \langle S_{\mathrm{T,R}} \rangle / \epsilon, \tag{6}$$

which in essence is the derivative of $\langle S_R \rangle$ or $\langle S_T \rangle$ in the q "direction".

With little algebra, we can show that the first order perturbations of the t-step entropies $\langle S_R \rangle$ and $\langle S_T \rangle$ are

$$\begin{aligned}\Delta\langle S_T\rangle &= -\epsilon \log \epsilon \sum_{i,j} \mathbb{1}_T^c P_j^{(0)} \eta_{ji}^{(t)} \\ -\epsilon \sum_{i,j} P_j^{(0)} \eta_{ji}^{(t)} &\left[\mathbb{1}_T (1 + \log(T^t)_{ji}) + \mathbb{1}_T^c \log \eta_{ji}^{(t)}\right]\end{aligned} \tag{7}$$

$$\begin{aligned}\Delta\langle S_R\rangle &= -\epsilon \log \epsilon \sum_{i,j} \mathbb{1}_T^c P_j^{(0)} \eta_{ji}^{(t)} \\ -\epsilon \sum_{i,j} P_j^{(0)} \eta_{ji}^{(t)} &\left[\log[(T^t)_{ji}/P_i^{(t)}] + \mathbb{1}_T^c \left(\log[\eta_{ji}^{(t)}/P_i^{(t)}] - 1\right)\right]\end{aligned} \tag{8}$$

The Kronecker functions $\mathbb{1}_T$ and $\mathbb{1}_T^c$ which implicitly depend on the indices i, j, and the time, t, are defined to be $\mathbb{1}_T = 0$ if $(T^t)_{ji} = 0$ and to equal 1 otherwise, and $\mathbb{1}_T^c = 1 - \mathbb{1}_T$.

While these equations ostensibly require a summation over all paths of the system, path integration in this setting is simply matrix multiplication, e.g., $\mathbf{T}^t$. This allows the calculation to be easily defined, and acomplished in polynomial time, $\mathcal{O}(n^3 \log t)$. The sums over states in Eqs. (7) and (8) are also of polynomial complexity, $\mathcal{O}(n^2)$.

As we see, the $\epsilon \log \epsilon$ terms can cause the limit, Eq. (6), to diverge, causing a sharp, singular change in entropy generation. This is expected. The divergence will happen only when the perturbation enables a path between two states where there was none. This is because $(T^t)_{ji} = 0$ only if i could not be reached from j in t steps, but if this is still true after perturbation, the η_{ji} term will be zero.

On the other hand, if the perturbation does not enable a path between two isolated states, but preserves the topology of the transition matrix, then Eq. (7) simplifies considerably; the divergent $\mathbb{1}^c$ terms vanish, and we take the limit,

$$\begin{aligned}\delta\langle S_T\rangle &= -\sum_{i,j} P_j^{(0)} \eta_{ji}^{(t)} [1 + \log(T^t)_{ji}] \\ \delta\langle S_R\rangle &= -\sum_{i,j} P_j^{(0)} \eta_{ji}^{(t)} \log[(T^t)_{ji}/P_i^{(t)}]\end{aligned} \tag{9}$$

Having established a very general theoretical framework, we now implement these ideas on two broad classes of stochastic systems for which the structure of the perturbing matrix q is specified further. We first consider random transition matrices drawn from a matrix ensemble. Second, we study a physical application—we enhance the predictability and retrodictability of thermalizing quantum mechanical systems by means of an external potential.

Improving the inferribility of Markov processes. We start by studying a general class of perturbations that can be applied to an arbitrary Markov process, and evaluate the associated entropy gradient, which can be thought as the direction in matrix space that locally changes $\langle S_R \rangle$ or $\langle S_T \rangle$ the most (Fig. 1). As we climb up or down the entropy gradient, we show how the transition matrix evolves (Fig. 2).

We consider a family of perturbations that vary the relative strength of any transition rate. This involves changing one element in the transition matrix while reallocating the difference to the remaining nonzero rates so that the total probability remains normalized. In other words,

$$\Delta_{\beta\alpha}^{(\epsilon)} T_{ji} = T_{ji} + \epsilon \cdot \mathbb{1}_{j\beta} (\mathbb{1}_{i\alpha} - T_{\beta i}). \tag{10}$$

To first order in ϵ, this is the same as adding ϵ to the (i, j) element, and then dividing the row by $1 + \epsilon$ to normalize, so it is a natural choice for a perturbation operator. It also obeys $\Delta^{(\epsilon)}\Delta^{(-\epsilon)}\mathbf{T} = \mathbb{1} + \mathcal{O}(\epsilon^2)$. We define the perturbation acting on a zero element to be zero if ϵ<0 since elements of the transition matrix must be non-negative. From Eqs. (10) and (4), we obtain the perturbed matrices and perturbed $\langle S_R \rangle$, $\langle S_T \rangle$.

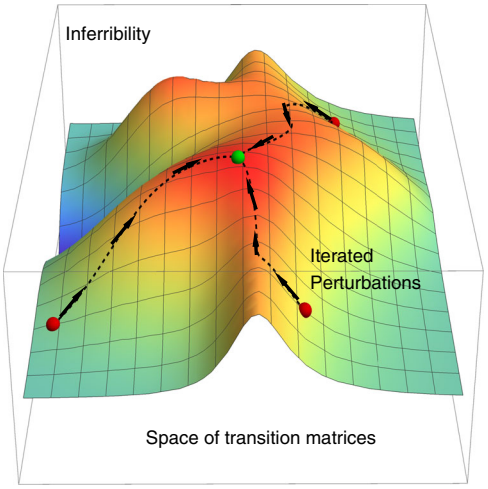

Fig. 1 Ascending the space of transition matrices to maximize predictability and retrodictability. Each point in the space of Markov transition matrices, represented by the x, y plane has an associated predictive and retrodictive entropy. Equation (11) allows us to find the direction in network space—parameterized by the transition rates T_{ji}—in which entropy locally increases (or decreases) the most. Perturbations can then be applied to move the network in that direction, leading to a system that is more susceptible to inference. Red dots represent different starting networks which climb along the black paths, via gradient ascent, to an entropy maximum, represented by a green dot

To study the effect of successive perturbations of the form Eq. (10), we carry out a gradient ascent algorithm in matrix space. At each iteration, we change the transition rates infinitesimally to maximally increase or decrease retrodiction or prediction entropy. We parameterize the gradient ascent by L^2 distance in matrix space, i.e., $\mathrm{d}(A, B) = \|A - B\|_2 = \left[\sum_{ij} (A_{ji} - B_{ji})^2\right]^{1/2}$.

In a gradient descent algorithm, one descends over a function $f(\mathbf{r})$ over a parameter t (time) by solving $\dot{\mathbf{r}} = \nabla f(\mathbf{r})/\|\nabla f(\mathbf{r})\|$, where the normalization ensures that $\|\partial_t \mathbf{r}(t)\| = 1$, so the total distance of the path $\mathbf{r}(t)$ just t.

Similarly, we define our gradients to be either $\Delta_{ji}\langle S_R(\mathbf{T})\rangle$ or $\Delta_{ji}\langle S_T(\mathbf{T})\rangle$, depending on whether we are optimizing retrodiction or prediction. We parameterize our path, $\mathbf{T}(\lambda)$, so that the total distance of the path (in L^2 matrix space) is λ,

$$\dot{T}_{ji}(\lambda) = \lim_{\epsilon \to 0} \Delta_{ji}^{(\epsilon)} \langle S_{T,R}(T_{ji})\rangle \Big/ \left\| \Delta_{ji}^{(\epsilon)} \langle S_{T,R}(T_{ji})\rangle \right\|. \tag{11}$$

Since we carry out this scheme numerically, using a finite difference method, it does not matter if the limit in Eq. (11) exists. In these cases, our numerical scheme returns large, finite jumps.

To illustrate our formalism in action, we solve Eq. (11) for a particular example system: diffusion taking place on a directed spatial random network. We build a spatial network, such that neighboring nodes are placed at regular intervals on a circle, and are also cross connected with probability $P(S_{ji} = 1) = \mathrm{e}^{-\beta d_{ij}}$, that decay with distance d_{ij}[42]. The transition matrix $\mathbf{T}$ is obtained by normalizing the rows of $\mathbf{S}$. For our prior, we use a uniform prior over all states.

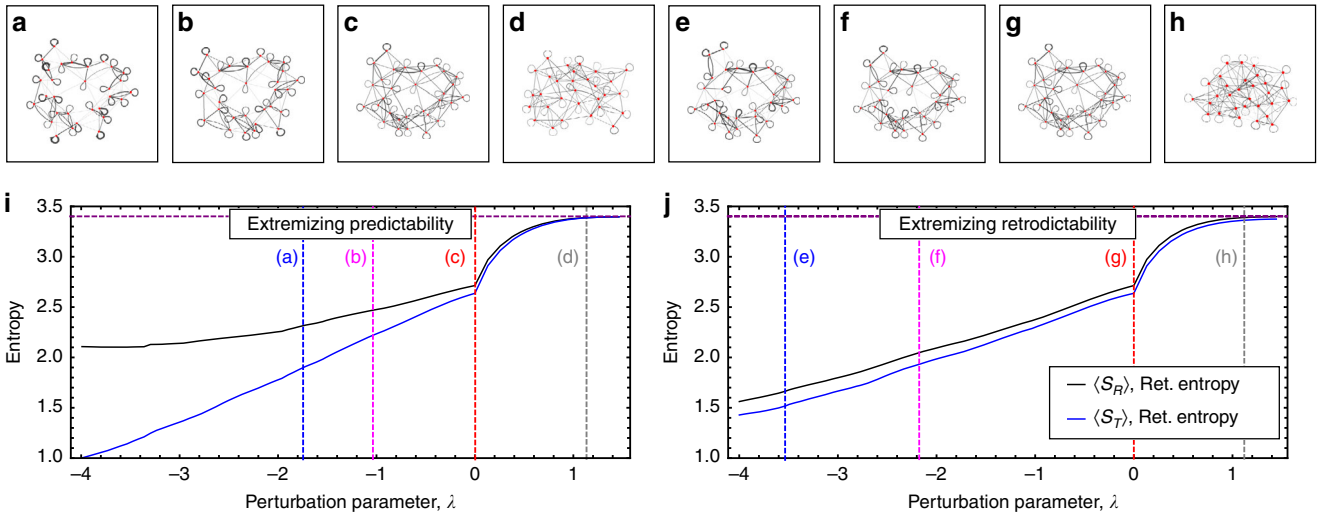

Fig. 2 Entropy extremization of a Markov process. Entropy during the evolution of a Markov network according to the extremization procedures, Eq. (11). The parameter λ corresponds to "how many times" the perturbation operator has been applied—it is the integrated L^2 distance of how far along the gradient curve, ∇S, we have pushed the transition network. The graphs in panels (**a**)-(**h**) are pictorial representations of the Markov transition matrices. The points in the evolution that we sample graphs from are marked with lines and the letter corresponding to the panels of graphs. The entropy curves, $\langle S_T \rangle$ and $\langle S_R \rangle$ correspond to how easy it is to predict the final state or retrodict the initial state of the Markov process. The network is a random geometric network, with adjacency matrix, **S**, picked from an ensemble $P(S_{ji} = 1) = e^{-\beta d_{ij}}$, where d_{ij} is the distance between i and j in a circular metric (where node n is adjacent to node 1). This is turned into a discrete diffusion matrix, **T**, by normalizing the rows of **S**. Here, $\beta = 0.5$ and $n = 30$ states. We optimize our entropies for a $t = 3$ step process. The purple line along the top marks the maximum possible entropy. **a**–**d** The graphs corresponding to 20%, 10%, 0% (original network), and −10% inference improvement, respectively, when extremizing $\langle S_T \rangle$. **e**–**h** The graphs corresponding to 20%, 10%, 0% (original network), and −9% inference improvement, respectively, when extremizing $\langle S_R \rangle$. **i** How the entropies change as we extremize $\langle S_T \rangle$, and four samples of transition probability networks. **j** How the entropies change as we extremize $\langle S_R \rangle$

An example is shown in Fig. 2, where the 3-step ($t = 3$) predictability and retrodictability change as the transition matrix is perturbed iteratively.

Snapshots of the perturbed matrix, for different λ values, when predictability is extremized can be seen in Fig. 2a–d, represented as networks with edge thicknesses proportional to the transition rates. Similarly, sample transition network when retrodictability is optimized can be seen in Fig. 2e–h. The behavior of prediction and retrodiction entropy as λ is varied can be seen in Fig. 2i, j, and dotted lines mark the λ values corresponding to the networks in Fig. 2a–h. We observe that inferential success can be improved up to 10–20% with only minor changes in the network structure. This will be quantified in further detail below.

We now interpret our results to ensure that our theoretical framework makes qualitative sense and works as expected. First, we observe that perturbations that maximize both $\langle S_T \rangle$ and $\langle S_R \rangle$ displace the transition matrix toward the same point: in both cases **T** evolves to a point where $(T^t)_{ji} = {}_{pi}$, a probability vector. In other words the probability of transition does not depend on what state the system is currently in. Taking the 3rd power of the **T** matrix for large values of λ reveals that this is indeed the case, although, of course, **T** itself can retain some complex structure. As expected, when a system moves from any state to any other state with equal likelihood, it is most difficult to infer its past or future.

In contrast, minimizing entropy produces two very different transition matrices, for $\lambda \ll 0$, depending on the type of entropy we minimize. The global minima of the prediction entropy are transition matrices in which all probability in each connected component flows toward a single node, reachable in t steps. A process where the initial state uniquely determines the final state is indeed trivial to predict. As we increase the number of time steps over which we optimize predictability, the number of "layers" of the transition network increases (Fig. 3a, b).

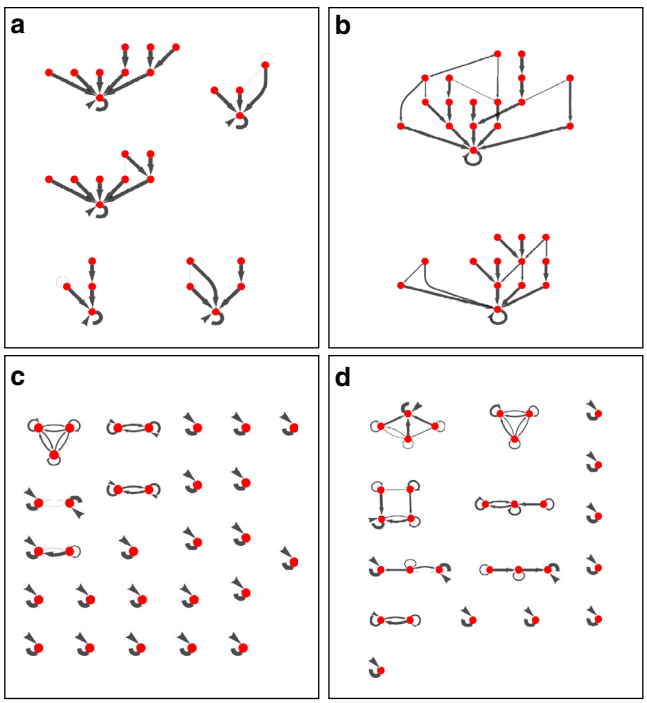

Fig. 3 Extremal networks. We start with a random transition network and show the structure of the transition matrix as it undergoes large amounts of optimization. We optimize predictability to the extent that the final state, given the initial state, could be correctly guessed on the first try 99% of the time (panels (**a**), (**b**)). We optimize retrodictability to the extent that given the final state, the initial state could be correctly guessed on the first try 75% (panels (**c**), (**d**)) of the time. Panels (**a**, **c**) optimize entropy for $s = 3$ step processes, whereas panels (**b**, **d**) do so for $s = 6$ step processes. Panels (**a**, **b**) optimize $\langle S_T \rangle$, while panels (**c**, **d**) optimize $\langle S_R \rangle$

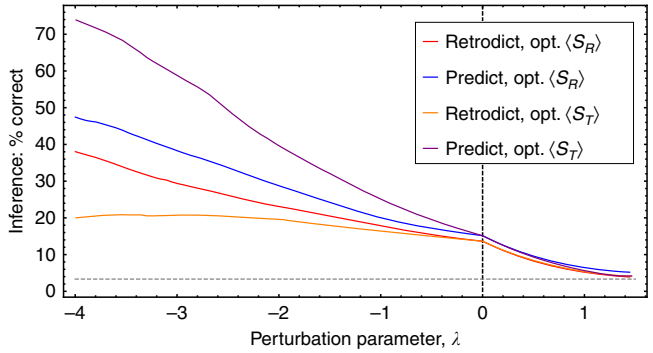

Fig. 4 Performance in predicting initial or final states. The performance of prediction and retrodiction on evolving random Markov transition networks. The four cases plotted are either correct inferences of the initial state (retrodiction) or correct inferences of the final state (prediction) while either optimizing $\langle S_R \rangle$ or optimizing $\langle S_T \rangle$. As a baseline, making random guesses, the strategy would obtain the initial or final state correctly 3.3% of the time (since there are 30 states). This baseline is depicted as a dashed gray line

On the other hand, minimizing retrodiction entropy tends to eliminate branches and fragmenting the network into linear chains (including isolated nodes). The start of this process can be seen in Fig. 2e. When the splitting is complete, probability flows through these fragments unidirectionally, thus retrodiction involves nothing more than tracing back a linear path (Fig. 3c, d).

This also explains why $\langle S_R \rangle$ tends to stay the same in the $\lambda < 0$ direction when minimizing $\langle S_T \rangle$. If $S_t = \langle S_T \rangle = 0$, then Eq. (2) implies $\langle S_R \rangle = S_0$, which is the maximum possible value for $\langle S_R \rangle$. This can also be understood intuitively—if when a final measurement is made, the system is always found to be in a unique accumulating state, this yields no information about what state the system started in. If, however, the minimal $\langle S_T \rangle$ network instead has multiple connected components and collector nodes, $\{k_j\}$, then there can be a decrease in $\langle S_R \rangle$ since R_{k_0}, R_{k_1},… are different distributions.

So far, we have only extremized entropy, but have not shown that this leads to a significant difference in our ability to infer the past or future. We will do so by reporting how often, on average, we can identify the correct initial (final) state of the system, given the final (initial) state. Note that this metric is not extensive; with increasing number of states, the probability mass for even the "best guess" approaches zero. Nevertheless, we will adopt this difficult metric for ourselves. For both predicting the final state and retrodicting the initial state, we perform a maximum likelihood inference; we pick the state with highest probability to be our guess, conditioned on the observed final or initial state. From the transition probability, T_{ji}, and retrodiction probability, R_{ji}, we can calculate the probability that our guess at the initial or final state will be correct (cf. "inference performance" in the Methods section).

We plot how inference performance changes as we manipulate transition rates in Fig. 4. The transition matrices we did our test is the same as those shown in Fig. 2. The success rate of predicting final states and retrodicting initial states while optimizing either $\langle S_R \rangle$ or $\langle S_T \rangle$ is plotted. Since there are 30 states in our network, the baseline accuracy is $1/30 = 3.3\%$, which is marked with a dashed gray line. Our success rate aligns well with the entropy in Fig. 2i. If we were to continue to larger values of λ, we reach almost 100% accuracy when we minimize $\langle S_T \rangle$ or $\langle S_R \rangle$.

The improvement in retrodictability always lags behind predictability. This is because $\langle S_R \rangle$ must be greater than $\langle S_T \rangle$, as per Eq. (2).

Table 1 Matrix retrodictability and structure

λ	Optimize	Δ%	L^2 dist.	L^1 dist.	$\Delta \geq 0.1$	Δ_{max}
+0.49	$\langle S_R \rangle$	−5.29%	0.49	8.96	4	0.12
−1.18	$\langle S_R \rangle$	+5.12%	1.03	9.69	39	0.23
−2.13	$\langle S_R \rangle$	+10.2%	1.63	16.2	86	0.31
+0.49	$\langle S_T \rangle$	−5.72%	0.49	8.81	4	0.13
−0.556	$\langle S_T \rangle$	+4.92%	0.54	4.34	8	0.16
−1.03	$\langle S_T \rangle$	+10.3%	0.91	7.81	23	0.22

A summary of how much the initial matrix must be changed to vary the retrodiction success. The columns are perturbation parameter (λ), what type of entropy was optimized (retrodiction or prediction, "optimize"), change in performance (retrodiction or prediction, corresponding to what was optimized) from the original matrix (Δ%), the L^2 distance from the original matrix, the L^1 distance from the original matrix, the number of transition probabilities that were changed more than 0.1 ($\Delta \geq 0.1$), and the maximum change of any transition probability. The original matrix had 126 nonzero entries and had retrodictability/predictability successes of 13.6 and 15.1%. There are 900 transition that can be modified

Naturally, descending an entropy landscape all the way returns transition matrices with trivial structure and dynamics. In our diffusion example, one could have guessed from the beginning, that a network with only inward branches, or one with disconnected linear chains, would be much more predictable than an all-to-all network with equally distributed weights. However, our formulation is useful not because it eventually transforms every network into a trivial network, but because it provides the steepest direction toward a trivial network. Second, our formulation is useful because, among many trivial networks, it moves us toward the direction of the closest one. Thus, we must determine the effectiveness of small pertubations, far before the system turns into a trivial one.

We find indeed, that significant differences to inferential success can be made with relatively small changes to the transition matrix. Table 1 quantifies how much the transition matrix has been modified, versus how much our retrodictive (top three rows) and predictive (bottom three rows) success have improved. For example, the fifth row shows that if we would like to be spot-on correct in predicting the final state of a stochastic process with 30 states and 900 transitions, our success rate can be improved by ~5% by modifying only 8 out of 900 transition rates by more than 0.1, with none being larger than 0.2. The cumulative change in all transition rates for this perturbation totals to 4.34, an equivalent of adding four edges. The changes required to improve our success rate by 10% are not much larger (Table 1).

As a final point of interest, we see that for all the ±5% in Table 1, λ and the L^2 distance are almost identical. This means that to get from the initial matrix to the perturbed matrices, one could follow the gradient calculated at the initial matrix in a straight line—the path is roughly straight in matrix space for at least that distance.

Improving the inferribility of quantum systems via external fields. In a physically realistic scenario, it is unlikely to have full control over individual transitions. An experimentalist can only tune physical parameters, such as external fields or temperature, which influence the transition matrix indirectly. Furthermore, it is often not practical to vary physical parameters by arbitrarily large amounts. Thus ideally we should improve predictability and retrodictability optimally, while only applying small fields.

To meet these goals, we consider a class of quantum systems in or out of equilibrium with a thermal bath. These systems are fully characterized by eigenstates ψ_1, …, ψ_n with energies E_1, …, E_n undergoing Metropolis–Hastings dynamics[43] where a system attempts to transition to an energy level above or below with equal probability; an attempt to decay always succeeds, while an

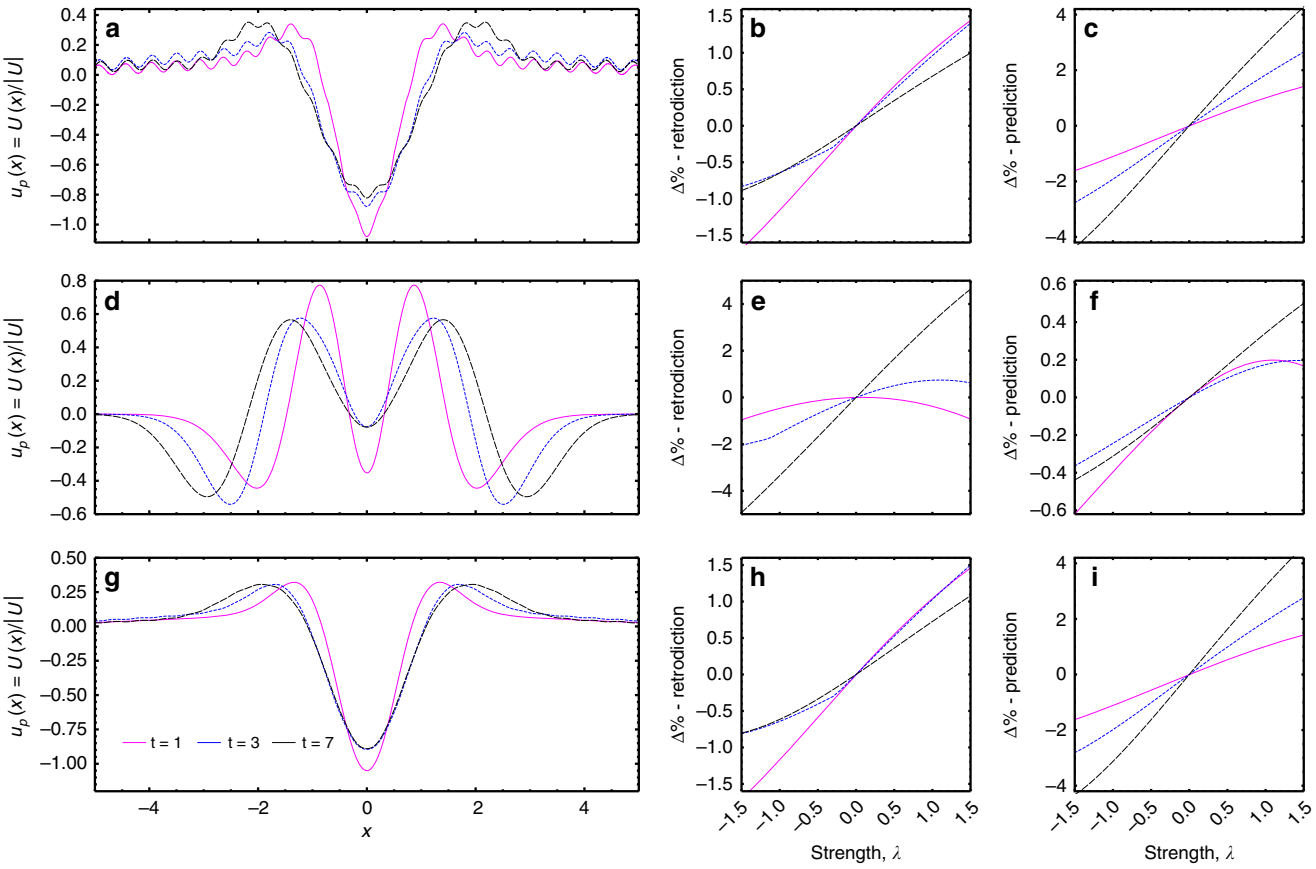

Fig. 5 The external fields and performance checks. We take $\hbar = \omega = m = 1$ and plot perturbations that minimize $\langle S_R \rangle$ or $\langle S_T \rangle$ for the quantum harmonic oscillator for processes taking $t = 1, 3, 7$ time steps, Eq. (18). The potentials, $U(x)$, which are (negatives of) the solutions to Eq. (18), are normalized by their L^2 norm. We plot this normalized potential, $u_p(x) = U(x)/\|U\|$. The normalized potential can be added to the system as a perturbation with different strengths, λ, i.e., $\hat{H}_{\rm osc} \to \hat{H}_{\rm osc} + \lambda u_p(x)$. Along with the normalized perturbing potential, we plot the change in inference success, Δ% (what percentage of the time the initial or final state of the system can be predicted or retrodicted) vs. strength of the applied perturbation. Locally, this curve should have positive slope. **a**–**c** A high temperature ($T = 10$) equilibrium system is quenched to a low temperature ($T = 1$) system. These potentials extremize $\langle S_R \rangle$. **d**–**f** A low temperature ($T = 1$) system quenched to a high temperature ($T = 10$) system. Note the large scale shape of the potential, panel (**d**), is similar to that of panel (**a**). These potentials extremize both $\langle S_R \rangle$ and $\langle S_T \rangle$. **g**–**i** A high temperature ($T = 10$) equilibrium system is quenched to a low temperature ($T = 1$) system. These potentials extremize $\langle S_T \rangle$

attempt to excite succeeds with probability $\exp[-\beta(E_{k+1} - E_k)]$.

$$T_{k,j} = \begin{cases} \frac{1}{2}\exp[-\beta(E_{k+1} - E_k)] & j = k+1 \\ \frac{1}{2}(1 - \exp[-\beta(E_{k+1} - E_k)]) & j = k \\ \frac{1}{2} & j = k-1 \\ 0 & |j-k|>1 \end{cases} \quad (12)$$

Furthermore we assume that the ground state E_0 cannot decay, and the highest state E_n is unexcitable. For the regime of validity of Markovian descriptions of thermalized quantum systems, we refer to refs. [44,45].

We now determine the effects of a small perturbing potential $v(x)$. The perturbation will shift the energy levels, which changes the transition matrix, which in turn changes the average prediction and retrodiction entropies of the system. Our goal is to identify what perturbing potential would maximally change these entropies. Since we are concerned with the first order variation in entropy, it will suffice to also use first order perturbation theory to calculate energy shifts.

The perturbed k-th energy level is $E_k = E_k^{(0)} + \epsilon \cdot \delta E_k$. When the perturbation is applied the exponential terms in $\mathbf{T}$ change as

$$e^{-\beta(E_{k+1}-E_k)} \to e^{-\beta(E_{k+1}-E_k)-\epsilon\beta(\delta E_{k+1}-\delta E_k)} = [1 - \epsilon\beta(\delta E_k - \delta E_{k-1})]e^{-\beta(E_k - E_{k-1})} + \mathcal{O}(\epsilon^2).$$

From this, we can find our first order change $T'_{ji} = T_{ji} + \epsilon q_{ji}$ in terms of the change in energy levels, δE_k,

$$q_{kj} = -\beta(\delta E_{k+1} - \delta E_k)\exp[-\beta(E_{k+1} - E_k)] \cdot S_{kj}$$
$$S_{kj} = \mathbb{1}_{j,k+1} - \mathbb{1}_{j,k} = \begin{cases} +1 & j = k+1 \\ -1 & j = k \\ 0 & j \neq k, j \neq k+1 \end{cases}. \quad (13)$$

Now we will write the prediction and retrodiction entropy $\delta\langle S_{T,R}\rangle$ variations as a functional of a perturbing potential, and then use calculus of variations to obtain the extremizing potential. For clarity, we will derive our equations in one dimension; the generalization to higher dimensions is straightforward.

We partition the spatial domain, Ω, into N intervals, $[x_i, x_{i+1})$, of width Δx and let our potential be a piecewise constant function of the form $v(x) = \sum_{i=0}^{N-1} v_i \mathbb{1}_{x\in[x_i,x_{i+1})}$. As $N \to \infty$, the first order

change in the k-th energy level is

$$\delta E_k = \langle \psi_k | v | \psi_k \rangle \sim \sum_{i=0}^{N-1} v_i |\psi(x_i)|^2 \Delta x \tag{14}$$

since $\int_{x_i}^{x_{i+1}} v_i |\psi(x)|^2 \sim v_i |\psi(x_i)|^2 \Delta x$. We substitute the δEs, Eq. (14), into Eq. (13) to get the q matrix,

$$\begin{aligned} q_{kj} &= \sum_{i=0}^{N-1} v_i \beta \left[|\psi_k(x_i)|^2 - |\psi_{k+1}(x_i)|^2 \right] \mathrm{e}^{-\beta(E_{k+1}-E_k)} S_{kj} \Delta x \\ &\equiv \sum_{i=0}^{N-1} v_i q_{kj}(x_i) \Delta x \to \int_\Omega v(x) \tilde{q}_{kj}(x) dx \end{aligned} \tag{15}$$

$$\tilde{q}_{kj}(x) \equiv \beta \left(|\psi_k(x)|^2 - |\psi_{k+1}(x)|^2 \right) \mathrm{e}^{\beta(E_{k+1}-E_k)} \cdot S_{kj} \tag{16}$$

we substitute this in into Eq. (5) to get

$$\eta_{ji}^{(t)} = \int_\Omega dx v(x) \sum_{k=0}^{t} (\mathbf{T}^k \tilde{\mathbf{q}}(x) \mathbf{T}^{t-k-1})_{ji} \equiv \int_\Omega dx v(x) \tilde{\eta}_{ji}^{(t)}(x)$$

$$\tilde{\eta}_{ji}^{(t)}(x) \equiv \sum_{k=0}^{t} (\mathbf{T}^k \tilde{\mathbf{q}}(x) \mathbf{T}^{t-k-1})_{ji}$$

and therefore,

$$\delta \langle S_{\mathrm{T,R}} \rangle [v] = -\int_\Omega dx v(x) \frac{\delta^2 \langle S_{\mathrm{T,R}} \rangle}{\delta x \delta v} \tag{17}$$

where $\delta^2 \langle S_{\mathrm{T,R}} \rangle / \delta x \delta v$ is Eq. (9) with $\tilde{\eta}_{ji}^{(t)}(x)$ substituted in for $\eta_{ji}^{(t)}$.

Last, we ensure the smallness of the perturbation by introducing a penalty functional, $C[v] = \frac{1}{2} \gamma \int v(x)^2 dx$ and ask what potential $v(x)$ extremizes

$$F_{\mathrm{T,R}} = \delta \langle S_{\mathrm{T,R}} \rangle - C = \int_\Omega \left(v(x) \frac{\delta^2 \langle S_{\mathrm{T,R}} \rangle}{\delta x \delta v} - \frac{1}{2} \gamma v(x)^2 \right) dx.$$

We take a variational derivative with respect to $v(x)$ and set it to zero to obtain the extremizing potential,

$$v_{\mathrm{T,R}}(x) = -\frac{1}{\gamma} \frac{\delta^2 \langle S_{\mathrm{T,R}} \rangle}{\delta x \delta v}. \tag{18}$$

This $v_{\mathrm{T,R}}$ is the external potential that extremizes the gradient of entropy minus the penalty functional.

Improving inferribility for a thermalizing quantum oscillator. We can now ask what perturbing external field should be applied a quantum harmonic oscillator that is in the process of warming up or cooling down, in order to improve its predictability or retrodictability. For this system $V(x) = \frac{1}{2} m \omega^2 x^2$, and $E_k = (k + \frac{1}{2}) \hbar \omega$. The stationary eigenfunctions are $\psi_k(x) = \frac{1}{\sqrt{2^k k!}} \pi^{-1/4} \exp\left(-\frac{x^2}{2}\right) H_k(x)$ where H_k is the k-th Hermite polynomial, $H_k(x) = (-1)^k \mathrm{e}^{x^2} \frac{d^k}{dx^k} \mathrm{e}^{-x^2}$. For concreteness, we also have to choose a prior distribution on states. We choose the prior distribution to be an equilibrium distribution at a (possibly different) temperature, $P_k \propto \mathrm{e}^{-\beta_2 E_k}$. We truncate the transition matrix at an energy $E_n \gg 1/\beta_1$, $1/\beta_2$ so that edge effects are negligible. We take $m = \hbar = \omega = 1$, and choose U to be the negative of Eq. (18) so that adding them to $V(x)$ decreases the corresponding entropy, and increases inference performance.

The initial and final temperatures determine the flow of probability. The equilibrium distribution at a high temperature has much more probability mass at higher energy states than an equilibrium distribution at a low temperature, so if we start with a high temperature and quench to a low temperature, there will tend to be a flow of probability from high states to low states. The opposite will happen when we quench from a low to a high temperature. We use $T = 1$ as the low temperature, and $T = 10$ as the high temperature.

To ensure that each perturbing potential, $U(x)$, actually increase or decrease predictability/retrodictability (depending on whether we add or subtract it from $V(x)$), we calculate the "Δ%" for retrodiction and prediction: the percent difference in how often we can correctly guess the initial or final state, upon perturbing the system. The performance is obtained similarly to that in Fig. 4 (cf. Methods section). The perturbation potential is normalized to $u_p(x) = U(x)/\|U\|$ so that the L^2 norm of $u_p(x)$ is 1, and the strength, λ, with which u_p is applied is varied so that the total potential is $V(x) + \lambda u_p(x)$.

Figure 5 shows some extremizing potentials for a system that was at one temperature, and is then suddenly quenched to a different temperature. Figure 5a shows optimal potentials for a system quenched from a high temperature to a low temperature while optimizing $\langle S_{\mathrm{R}} \rangle$ for $t = 1$, 3, and 5 time steps. Figure 5b, c shows the change in inference success as the potential is applied at varying strengths, λ. Figure 5d–f shows the same quantities (extremizing potential and change in inference) for a system at a low temperature quenched to a high temperature, while optimizing $\langle S_{\mathrm{R}} \rangle$. This potential is also optimal for $\langle S_{\mathrm{T}} \rangle$. Finally, Fig. 5g–i shows a high temperature system quenched to a low temperature while optimizing $\langle S_{\mathrm{T}} \rangle$.

To quantify how significantly the perturbations change the quantum system, we keep track of the L^1 difference in eigenvalue spacing, i.e., $\mathcal{S} \equiv \sum_{k=1}^{30} |E_k' - E_{k-1}' - \hbar\omega|$. The largest values $\mathcal{S}$ achieves for any potential and applied strength shown in Fig. 5 is ~1.2. In other words, we can get few percent change in success rate by introducing a change to all energy levels that amounts to one level spacing. Note that this is a single step perturbation along a single direction, rather than an iterated one.

This example illustrates how to combine real, physical, continuous quantities, such as perturbation potentials, with the more abstract formalism of evaluating the entropy of Markov transition matrices with discrete states. The general procedure we outlined in this section can also be applied to other thermal systems, quantum or otherwise.

Discussion

We developed a formalism to describe exactly how predictability and retrodictability changes in response to small changes in a transition matrix, and used it to descend entropy landscapes to optimally improve the accuracy with which the past or future of a stochastic system can be inferred. Our main results are the equations relating perturbations of Markov processes to the change in average entropy and retrodiction entropy of the system, Eqs. (4), (7), and (9).

We specifically focused on Markov processes, not only because it yields to mathematical analysis, but also because many important processes in physical, biological and social sciences are Markovian. That being said, the general principle outlined here can also be used in systems with memory, or in other inference problems such as the determination of unknown boundary conditions, system parameters, or driving forces.

As examples of manipulating predictability and retrodictability, we studied two specific types of perturbations, Eqs. (10) and (13), and used these to study how certain types of transition matrices evolve as they flow along the trajectory of maximal increase in retrodictability and predictability. We found that the transition networks tend to cull their connections and split into cycles and chains when we try to minimize retrodiction entropy. Conversely, the transition networks become fully connected when we attempt

to maximize either inferential entropy. If one does not have full control over transition rates, one can steer a system toward the direction of either extreme by a small amount. Finally, as a physical example, we studied how to find the perturbing potential that extremally changes the predictability and retrodictability of a thermalizing quantum system.

Our formulas lead us to intuitive results such as the divergence of entropy generation when a path between two otherwise isolated states is enabled. However, they also lead us to less obvious conclusions, such as how predictability changes when retrodictability is optimized (and vice versa); or the shape of optimal potentials perturbing a thermalizing quantum system.

Our basic equations, Eqs. (4), (7), and (9), are very generally applicable to any discrete-time Markov process. The type of transition matrix perturbations we chose to study, namely those in Eqs. (10) and (13) are natural and practical choices, but of course, they are not the only two possibilities. For example, an operator that takes two matrix elements $0 < T_{ja},\ T_{jb} < 1$ and "transfers" probability between them, changing them to $T_{ja} + \epsilon, T_{jb} - \epsilon$ would make an interesting future study.

In our work we observed an intriguing asymmetry between prediction and retrodiction. In particular, we observe that predictability is more easily improved than retrodictability. This a byproduct of how we set up our problem: We took the initial distribution, the probability vector $\mathbf{P}^{(0)}$, and the forward dynamics, $\mathbf{T}$, as givens, and found the probability, $\mathbf{P}^{(t)}$, via propagating $\mathbf{P}^{(0)}$ with $\mathbf{T}$. If we had done the opposite by picking the distribution $\mathbf{P}^{(t)}$ and the backwards dynamics, $\widetilde{\mathbf{T}}$, then we could find $\mathbf{P}^{(0)}$ to be the back-evolved distribution, then our results would reverse.

An experimenter only has control over the prior distribution at the current time, $\mathbf{P}^{(0)}$, but cannot in general decide what distribution she wants at an arbitrary future time, $\mathbf{P}^{(t)}$, and pick a $\mathbf{P}^{(0)}$ that results in a specified $\mathbf{P}^{(t)}$. The fact that we set up the problem so that $t = 0$ was the "controlled" time, and the state at the final time is the result of the choices made at $t = 0$ ultimately lead to the seeming emergence of an "arrow of time"[46].

Since our method makes changes to a system to extremize the average of a function over a set of trajectories, it could well be considered within the domain of stochastic control theory[47,48]. However, there are various elements in our approach that depart from classical stochastic control, which typically deals with problems of the form

$$dX_t = f(X_t, \nu(t); t) + \hat{\xi}_t$$
$$C(X_0, \nu) = \left\langle \phi(X_T) + \int_0^T R(X_t, \nu; t)\, dt \right\rangle_{P(\xi)}$$

where X_t is the system trajectory, $\hat{\xi}$ is a Weiner process, ν is a control parameter, C is a cost function, and ϕ and R are the target cost and some function that quantifies cost-of-control, cost-of-space, cost-of-dynamics, etc[49]. The goal is to find the $\tilde{\nu}$ that minimizes C.

One difference is that we do not restrict ourselves to a Weiner process, but allow any valid transition matrix. The control parameter, ν, could be the perturbation to the original transition matrix, or it could be some other external parameter which indirectly results in a change in the transition matrix, as in the thermalizing quantum oscillator example.

The second difference is the structure of our cost function. In our case, the cost is an average weighted over priors. For prediction entropy,

$$\langle S_{\mathrm{T}} \rangle = \langle C(X_0) \rangle_{P^{(0)}} = -\langle \langle \log P(X_T|X_0) \rangle_{P(\cdot|X_0)} \rangle_{P^{(0)}}.$$

For a delta function prior, this reduces to the standard control theory cost function, which depends on the initial condition of the system. For retrodiction entropy $S_R(X_T)$ the cost depends on the final state, and is then averaged over the posterior distribution of X_T,

$$\langle S_{\mathrm{R}} \rangle = \langle C(X_T) \rangle_{P^{(T)}} = -\langle \log R(X_0|X_T) \rangle_{R(\cdot|X_T)} \rangle_{P^{(T)}}.$$

The third difference is a philosophical one. Standard stochastic control aims to find a control protocol that is a global minimum of the cost function—one obtains the field ν such that $C[\nu + \delta\nu] = 0$ for all $\delta\nu$. In contrast, we look for the variation $\delta\nu$ such that $C[\delta\nu]$ is maximal, where C is $\langle S_{\mathrm{R}} \rangle$ or $\langle S_{\mathrm{T}} \rangle$. Our method descends entropy gradients in a space of system parameters, and is only guaranteed to be optimal locally. This could then be paired with a stochastic gradient descent algorithm or simulated annealing to find optima in a larger neighborhood. In passing, we note that for systems with a very large number of states, it would probably be computationally advantageous to use a stochastic algorithm even to compute the local gradient.

There is still plenty of room to make our framework more useful and general. Currently, we assume constant transition rates, and perturb the transition matrix at a single instant. However, transition rates can be time-dependent, in which case we would have to perturb the transition rates differently at different times. A second interesting avenue would be to further explore the costs associated with changing the transition probabilities. Another natural generalization is to extend the problem to continuous time.

Methods

Extremization of entropy. We started with a random geometric graph, $\mathbf{T}(\lambda = 0)$, from the ensemble described in the text, where nodes i and j are connected with probability $e^{-\beta d(i,j)}$. We used $n = 30$ node graphs, with $\beta = 0.5$. The extremization is done numerically and iteratively, as outlined in Eq. (11). The entropy was the entropy for a $t = 3$ step process, and we use a perturbation size $\epsilon = 0.05$, and step size $d\lambda = 0.05$.

At each step, the matrix of change in entropy (per ϵ) due to perturbation of an element is calculated, $S_{ji} = \frac{1}{\epsilon}\Delta_{ji}^{(\epsilon)} \langle S[\mathbf{T}(\lambda)] \rangle$, where the S in the angled brackets is whichever entropy we seek to extremize $\mathbf{T}$ over—either $\langle S_{\mathrm{R}} \rangle$ or $\langle S_{\mathrm{T}} \rangle$. To get the updated transition matrix, the j, i element of $\mathbf{T}$ is perturbed using the standard perturbation operator, Eq. (10), and strength $\epsilon' = d\lambda / \left\| S_{ji} \right\|$. The order that we apply these operators is irrelevant up to order $(\epsilon')^2$. The updated transition matrix is then the result of applying all the perturbation operators, one for each element of $\mathbf{T}$. At each step, the prediction and retrodiction entropy of the Markov process were calculated and saved, along with the actual matrix $T_{ji}(\lambda)$, for plotting purposes. The change in λ at each step is just the L^2 distance between the previous matrix and the new, perturbed matrix.

Inference performance. The inference performance can also be calculated analytically as long as we have the transition matrix, T_{ji}, and the prior, $P^{(0)}$, which we do to generate Figs. 4 and 5. Since we are guessing that the maximally likely state is the correct, the formulas for this are

$$C_{\mathrm{T}} = \sum_j P_j^{(0)} \max_i (T^t)_{ji}$$
$$C_{\mathrm{R}} = \sum_j P_j^{(t)} \max_i (R^t)_{ji}. \tag{19}$$

These formulas give us the expected fraction of times we correcly guess the final state given the initial state (C_{T}), or initial state given the final state (C_{R}). The expression $\max_i (T^t)_{ji}$ is the probability that you guess the the final state correctly given that the initial state is i, and the normalized sum simply averages your performance across all possible initial states. The C_{R} equation is analogous, simply substituting the retrodiction probability matrix for the transition matrix.

As expected, the performance obtained via random trials fits C_{T}, C_{R} almost exactly since we are using a large number of trials.

Thermalizing quantum harmonic oscillator. While it would be difficult to analytically solve Eqs. (18) and (9) to find the extremal change in potential, it is a simple matter to calculate it analytically.

For the harmonic oscillator $V(x) = \frac{1}{2} m\omega x^2$, and $E_k = (k + \frac{1}{2})\hbar\omega$. The stationary eigenfunctions are $\psi_k(x) = \frac{1}{\sqrt{2^k k!}} \pi^{-1/4} \exp\left(-\frac{x^2}{2}\right) H_k(x)$ where H_k is the k-th Hermite polynomial, $H_k(x) = (-1)^k e^{x^2} \frac{d^k}{dx^k} e^{-x^2}$. As mentioned in the text, we

choose the prior distribution to be an equilibrium distribution at a given temperature, $P_k \propto e^{-\beta_2 E_k}$. Since we can only store finite vectors on a computer, we only track the first $n = 30$ energy eigenstates, which is enough that the total probability mass (sum of Gibbs factors) the prior misses by truncation would only be <5% of the total probability mass. We take $m = \hbar = \omega = 1$ for simplicity. The perturbation matrix $\eta(x)$ can be calculated numerically—it is a high order (order 60) polynomial in x times e^{-x^2}—and substituted into Eq. (9) to get the (negative) extremal potential, $U(x)$. The potential is then normalized by the L^2 norm of U, $u_p(x) = U(x)/\|U\|$ where $\|U\| = \left(\int_{-\infty}^{\infty} U(x)^2 \mathbf{d}x\right)^{1/2}$.

Inference performance for quantum harmonic. We solve for the energy eigenvalues of the harmonic oscillator potential plus the perturbation potentials using the method of shooting. For each u_p, and for each strength, γ, we numerically solve Schrodinger's equation for the potential $\frac{1}{2}m\omega^2x^2 + \gamma u_p(x)$ at different energies, E_{trial}. We pick our shooting point to be far outside our region of interest, at $x = 15$, and evaluate whether the value of the numerical solution is positive or negative at the shooting point. Near an energy eigenvalue, the sign of ψ_{trial} will be (without loss of generality) <0 for energies a little below the true eigenvalue, and >0 for energies a little above the true eigenvalue. We use the bisection method of root finding to approximate the energy eigenvalue, with as much precision as we want. Our energy eigenvalues are correct up to 10^{-6}.

Once we have found the first n eigenvalues, we compute the transition matrix using Eq. (12), which is determined by the final temperature, and the prior distribution on states, $P_j^{(0)} = \mathrm{e}^{-\beta_1 E_j}/Z$, which is determined by the initial temperature ($Z = \sum_{j=1}^{n} \mathrm{e}^{-\beta_i E_j}$). We then calculate the average percentage of times the final state can be inferred given the initial state after $t = 1$, 3, 7 steps. We use Eq. (19) to do this.

Data availability

Both the data and the code used to create and analyze the data during the current study are available in the Github repository, https://github.com/nrupprecht/Retrodiction-Data/.

Received: 5 December 2018 Accepted: 2 May 2019
Published online: 19 June 2019

References

1. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis* 2 (Wiley, New York, 1958).
2. Le Cam, L. Maximum likelihood: an introduction. *Int. Stat. Rev.* **58,** 153–171 (1990).
3. Box, G. E. & Tiao, G. Bayesian Inference in Statistical Analysis (John Wiley & Sons, New York, 2011).
4. Turner, D. The functions of fossils: inference and explanation in functional morphology. *Stud. Hist. Philos. Sci. Part C: Stud. Hist. Philos. Biol. Biomed. Sci.* **31**, 193–212 (2000).
5. Slater, G. J., Harmon, L. J. & Alfaro, M. E. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evol.: Int. J. Org. Evol.* **66**, 3931–3944 (2012).
6. Gavryushkina, A., Welch, D., Stadler, T. & Drummond, A. J. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput. Biol.* **10**, e1003919 (2014).
7. Krauss, L. M. & Starkman, G. D. Life, the universe, and nothing: life and death in an ever-expanding universe. *Astrophys. J.* **531**, 22 (2000).
8. Ulanowicz, R. E. Increasing entropy: heat death or perpetual harmonies? *Int. J. Des. Nat. Ecodynamics* **4**, 83–96 (2009).
9. Frautschi, S. Entropy in an expanding universe. *Science* **217**, 593–599 (1982).
10. Baum, L. E. & Petrie, T. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.* **37**, 1554–1563 (1966).
11. Nasrabadi, N. M. Pattern recognition and machine learning. *J. Electron. imaging* **16**, 049901 (2007).
12. Fine, S., Singer, Y. & Tishby, N. The hierarchical hidden markov model: analysis and applications. *Mach. Learn.* **32**, 41–62 (1998).
13. Boyen, X. & Koller, D. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 33–42 (Morgan Kaufmann Publishers Inc., San Francisco, 1998).
14. Stevenson, I. H., Rebesco, J. M., Miller, L. E. & Körding, K. P. Inferring functional connections between neurons. *Curr. Opin. Neurobiol.* **18**, 582–588 (2008).
15. Nguyen, H. C., Zecchina, R. & Berg, J. Inverse statistical problems: from the inverse ising problem to data science. *Adv. Phys.* **66**, 197–261 (2017).
16. Ghonge, S. & Vural, D. C. Inferring network structure from cascades. *Phys. Rev. E* **96**, 012319 (2017).
17. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Series B* **36**, 192–225 (1974).
18. Cocco, S. & Monasson, R. Reconstructing a random potential from its random walks. *EPL (Europhys. Lett.)* **81**, 20002 (2007).
19. Iba, H. Inference of differential equation models by genetic programming. *Inf. Sci.* **178**, 4453–4468 (2008).
20. Gomez Rodriguez, M., Leskovec, J. & Krause, A. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1019–1028 (ACM, New York, 2010).
21. Lenglet, C., Deriche, R. & Faugeras, O. Inferring White Matter Geometry from Diffusion Tensor MRI: Application to Connectivity Mapping. *European Conference on Computer Vision*, 127–140 (Springer, Heidelberg, 2004).
22. Haas, K. R., Yang, H. & Chu, J.-W. Expectation-maximization of the potential of mean force and diffusion coefficient in langevin dynamics from single molecule fret data photon by photon. *J. Phys. Chem. B* **117**, 15591–15605 (2013).
23. Ghahramani, Z. & Hinton, G. E. Parameter estimation for linear dynamical systems. Tech. Rep., Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science (1996).
24. Lokhov, A. Y., Mézard, M., Ohta, H. & Zdeborová, L. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* **90**, 012801 (2014).
25. Altarelli, F., Braunstein, A., Dall'Asta, L., Ingrosso, A. & Zecchina, R. The patient-zero problem with noisy observations. *J. Stat. Mech.: Theory Exp.* **2014**, P10016 (2014).
26. Vural, D. C. Vural dc. when models interact with their subjects: the dynamics of model aware systems. *PLoS One* **6**, e20721 (2011).
27. Rupprecht, N. & Vural, D. C. Collective motion of predictive swarms. *PloS One* **12**, e0186785 (2017).
28. Crutchfield, J. P., Ellison, C. J. & Mahoney, J. R. Time's barbed arrow: irreversibility, crypticity, and stored information. *Phys. Rev. Lett.* **103**, 094101 (2009).
29. Ellison, C. J., Mahoney, J. R. & Crutchfield, J. P. Prediction, retrodiction, and the amount of information stored in the present. *J. Stat. Phys.* **136**, 1005 (2009).
30. Tatem, A. J., Rogers, D. J. & Hay, S. Global transport networks and infectious disease spread. *Adv. Parasitol.* **62**, 293–343 (2006).
31. Rupprecht, N. & Vural, D. C. Limits on inferring the past. *Phys. Rev. E* **97**, 062155 (2018).
32. Farid Golnaraghi, B. C. K. *Automatic Control Systems* (John Wiley & Sons, Hoboken, 1972).
33. Carnevale, G., Frisch, U. & Salmon, R. H theorems in statistical fluid dynamics. *J. Phys. A: Math. Gen.* **14**, 1701 (1981).
34. Ramshaw, J. D. H-theorems for the tsallis and renyi entropies. *Phys. Lett. A* **175**, 169–170 (1993).
35. Shiino, M. Free energies based on generalized entropies and H-theorems for nonlinear Fokker–Planck equations. *J. Math. Phys.* **42**, 2540–2553 (2001).
36. Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **5**, 3–55 (2001).
37. Lemons, D. S. & Langevin, P. *An Introduction to Stochastic Processes in Physics* (JHU Press, Baltimore, 2002).
38. Prinz, J.-H. et al. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
39. Urban, D. L. Modeling ecological processes across scales. *Ecology* **86**, 1996–2006 (2005).
40. Black, A. J. & McKane, A. J. Stochastic formulation of ecological models and their applications. *Trends Ecol. Evol.* **27**, 337–345 (2012).
41. Rohlf, K., Fraser, S. & Kapral, R. Reactive multiparticle collision dynamics. *Comput. Phys. Commun.* **179**, 132–139 (2008).
42. Barthélemy, M. Spatial networks. *Phys. Rep.* **499**, 1–101 (2011).
43. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
44. Gardiner, C., Zoller, P. & Zoller, P. *Quantum Noise: a Handbook of Markovian and Non-Markovian Quantum Stochastic Methods with Applications to Quantum Optics*, vol. 56 (Springer Science & Business Media, Heidelberg, 2004).
45. Kapral, R. Progress in the theory of mixed quantum-classical dynamics. *Annu. Rev. Phys. Chem.* **57**, 129–157 (2006).
46. Coveney, P. & Highfield, R. *The Arrow of Time: A Voyage Through Science to Solve Time's Greatest Mystery* (Fawcett Columbine, New York, 1992).
47. Åström, K. J. *Introduction to Stochastic Control Theory* (Academic Press, Inc., New York, 1970).
48. Forte, G. & Vural, D. C. Iterative control strategies for nonlinear systems. *Phys. Rev. E* **96**, 012102 (2017).

49. Chernyak, V. Y., Chertkov, M., Bierkens, J. & Kappen, H. J. Stochastic optimal control as non-equilibrium statistical mechanics: calculus of variations over density and current. *J. Phys. A: Math. Theor.* **47**, 022001 (2013).

Acknowledgements
This study was supported by National Science Foundation grants CBET-1805157.

Author contributions
N.R. and D.C.V. conceived the problem, interpreted the results, and wrote the paper. N.R. carried out the calculations and simulations.

Additional information
Competing interests: The authors declare no competing interests.